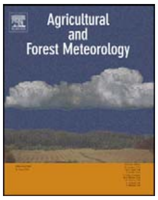
Agricultural and Forest Meteorology 149 (2009) 1597–1615

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet



The REFLEX project: Comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data

Andrew Fox[a], Mathew Williams[b,*], Andrew D. Richardson[c], David Cameron[d], Jeffrey H. Gove[e], Tristan Quaife[f], Daniel Ricciuto[g], Markus Reichstein[h], Enrico Tomelleri[h], Cathy M. Trudinger[i], Mark T. Van Wijk[j]

[a] School of Applied Maths, Centre for Terrestrial Carbon Dynamics, University of Sheffield, Sheffield, UK
[b] School of GeoSciences, Centre for Terrestrial Carbon Dynamics, University of Edinburgh, Edinburgh, UK
[c] Complex Systems Research Center, University of New Hampshire, Durham, NH, USA
[d] Centre for Ecology and Hydrology, Bush Estate, Penicuik, Midlothian, UK
[e] USDA Forest Service, Northern Research Station, Durham, NH, USA
[f] Centre for Terrestrial Carbon Dynamics, Department of Geography, UCL, London, UK
[g] Oak Ridge National Laboratory, Oak Ridge, TN, USA
[h] Max Planck Institute for Biogeochemistry, Jena, Germany
[i] CSIRO Marine and Atmospheric Research, Centre for Australian Weather and Climate Research, Aspendale, Victoria, Australia
[j] Wageningen University, Plant Sciences, Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 30 January 2009
Received in revised form 28 April 2009
Accepted 1 May 2009

Keywords:
Data assimilation
Metropolis
Carbon cycle
Ecosystem modelling
Monte Carlo
Kalman filter
Eddy covariance
REFLEX project
Parameter optimisation
Confidence intervals

ABSTRACT

We describe a model-data fusion (MDF) inter-comparison project (REFLEX), which compared various algorithms for estimating carbon (C) model parameters consistent with both measured carbon fluxes and states and a simple C model. Participants were provided with the model and with both synthetic net ecosystem exchange (NEE) of CO_2 and leaf area index (LAI) data, generated from the model with added noise, and observed NEE and LAI data from two eddy covariance sites. Participants endeavoured to estimate model parameters and states consistent with the model for all cases over the two years for which data were provided, and generate predictions for one additional year without observations. Nine participants contributed results using Metropolis algorithms, Kalman filters and a genetic algorithm. For the synthetic data case, parameter estimates compared well with the true values. The results of the analyses indicated that parameters linked directly to gross primary production (GPP) and ecosystem respiration, such as those related to foliage allocation and turnover, or temperature sensitivity of heterotrophic respiration, were best constrained and characterised. Poorly estimated parameters were those related to the allocation to and turnover of fine root/wood pools. Estimates of confidence intervals varied among algorithms, but several algorithms successfully located the true values of annual fluxes from synthetic experiments within relatively narrow 90% confidence intervals, achieving >80% success rate and mean NEE confidence intervals $<110\ \text{gC}\ \text{m}^{-2}\ \text{year}^{-1}$ for the synthetic case. Annual C flux estimates generated by participants generally agreed with gap-filling approaches using half-hourly data. The estimation of ecosystem respiration and GPP through MDF agreed well with outputs from partitioning studies using half-hourly data. Confidence limits on annual NEE increased by an average of 88% in the prediction year compared to the previous year, when data were available. Confidence intervals on annual NEE increased by 30% when observed data were used instead of synthetic data, reflecting and quantifying the addition of model error. Finally, our analyses indicated that incorporating additional constraints, using data on C pools (wood, soil and fine roots) would help to reduce uncertainties for model parameters poorly served by eddy covariance data.

1. Introduction

The land surface modelling community is increasingly seeking to test its terrestrial ecosystem models against the growing array of observations (Bonan, 2008). Such model-data comparison provides an opportunity to highlight areas in space or time of poor process representation, and to guide model improvement. A critical dataset to be used in evaluating ecosystem models is that of eddy covariance (EC) data (Baldocchi et al., 2001), which are collected from hundreds of sites worldwide, some over more than a decade. However, these data are associated with uncertainties and complications (Lasslop et al., 2008; Richardson et al., 2008). EC

[*] Corresponding author. Tel.: +44 131 667 8631; fax: +44 131 662 0478.
E-mail address: mat.williams@ed.ac.uk (M. Williams).

0168-1923/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.agrformet.2009.05.002

towers measure net ecosystem exchanges (NEE) of CO_2, meaning that the underlying processes of photosynthesis (GPP) and ecosystem respiration (R_e) that are represented in models are not directly measured during daytime (Desai et al., 2008).

A meaningful comparison between model and data is complicated by the need to assess and account for both model and observational errors. Thus, the probability of a model being correct should be assessed by taking into account observational uncertainties. Model uncertainty is also an important factor in any comparison with data. Models may be uncertain because of how they represent key processes, how initial conditions are set, or because their parameters are poorly determined. Separating these causes of uncertainty is important for guiding model development.

Model-data fusion (MDF) approaches, previously used mainly in hydrology and weather forecasting, are now being used more frequently by the terrestrial C cycle community (Raupach et al., 2005). MDF combines models with observations, taking account of model and observational uncertainties. In theory, MDF provides a means to cope with the problems arising from incomplete and noisy observational data, and uncertainty in model processes, initial states and parameters. MDF combines models with observations, and estimates of their uncertainties, to produce estimates of system dynamics with confidence intervals (Williams et al., 2005) and model parameterisations consistent with data. We refer to these outputs of MDF schemes as "analyses" hereafter.

The capabilities and weaknesses of the various existing MDF approaches remain poorly understood. One recent study, the OptIC experiment, used pseudo-data from a highly simplified test model with four parameters to compare parameter estimation methods (Trudinger et al., 2007). OptIC found different methods equally successful, but that the choice of the cost function (quantifying the model-data mismatch) caused the most variation in the estimated parameters. OptIC also demonstrated that the effort expended and experience of the user was a factor in successful solutions. However, OptIC did not use observed data of C fluxes, nor did it test state estimation or model forecast capabilities. With observed data, MDF is complicated by observational and model error and bias.

Here we describe the REgional FLux Estimation eXperiment. REFLEX is a model-data fusion inter-comparison project, aimed at comparing the strengths and weaknesses of various MDF algorithms for estimating parameters, fluxes and states of a C model. REFLEX participants used a mass balance C dynamics model that links C fluxes to changes in C stocks. The model generates NEE estimates based on its description of photosynthesis, autotrophic and heterotrophic respiration, all generated as functions of C stocks. These predictions were fused with either observed or synthetic daily NEE data. Unlike OptIC, real data were used in REFLEX, and the model employed is able to simulate GPP and R_e, and thus predict NEE, for direct comparison with eddy covariance data. The key question addressed here is: what are the confidence intervals on model parameters calibrated from EC data, and on model analyses and predictions of net C exchange and carbon stocks over multiple years? The experiment has an explicit focus on how different algorithms and protocols quantify the confidence intervals on parameter and state estimates, given the same C budget model and datasets.

What is novel in REFLEX is an explicit focus on comparing how an ensemble of MDF algorithms perform in terms of estimating C model states and parameters, and the uncertainties on these quantities. By using a single common model and both synthetic and observed EC datasets, and diagnostic and prognostic tests, we are able to generate insights into current capabilities for assessing and forecasting ecosystem NEE using model-data fusion.

2. Methods

In REFLEX, participants first used synthetic data and their choice of MDF algorithm to retrieve parameters and states consistent with a specified C model. Synthetic data were generated from the specified C model with a certain parametrization, with noise and gaps added to model outputs. The synthetic experiment dealt with observational and algorithmic error, and also user error including assumptions related to initial conditions and parameter priors. There was no model error or driver error. REFLEX participants then went on to fuse data from EC systems and local measurements of leaf area index (LAI) with the C model. This exercise introduced model and, to a lesser extent, driver error, because the model used does not perfectly describe forest ecosystem C fluxes, and because meteorological observations may contain small errors. Finally, REFLEX participants used the C model in a prognostic, rather than diagnostic, mode. One year of daily driver data was provided to produce forecasts of C dynamics, using parameters generated in the diagnoses, and the forecasts were tested against withheld data, both synthetic and observed (Table 1).

2.1. Model description

The requirements for the REFLEX C model included simplicity, a C mass balance, and vegetation and soil C pools with time constants covering days to decades. The model outputs had to include daily NEE and LAI. We selected the Data Assimilation Linked Ecosystem Carbon (DALEC) model (Williams et al., 2005), originally designed for evergreen forests, and a modified version (DALEC-D) for deciduous forests (Fig. 1). DALEC is a simple box model of carbon pools connected via fluxes running at a daily time-step. For the evergreen model there are five C pools representing foliage (C_f), woody stems and coarse roots (C_w), and fine roots (C_r) along with fresh leaf and fine root litter (C_{lit}) and soil organic matter and coarse woody debris (C_{som}). In the deciduous model there is an additional labile pool (C_{lab}) of stored C to support leaf flushing. The following assumptions were made to determine the fluxes between the C pools:

1. All C fixed during a day is expended either in autotrophic respiration or else allocated to one of the three plant tissue pools, C_f, C_w or C_r.
2. Autotrophic respiration is a constant fraction of the C fixed during a day (Waring et al., 1998).
3. Allocation fractions to vegetation pools are donor-controlled functions which have constant rate parameters.

Table 1
Experimental summary for REFLEX. The table shows for each experiment the input data, the source of the meteorological drivers, and the site codes. The first two experiments generated parameter estimates and estimates of model states (fluxes and pools of C), while the final two experiments were forecasts of model states only. Acronyms: DE – deciduous vegetation; EV – evergreen vegetation; SYN – synthetic data; EC – observed eddy covariance and LAI data.

Experiment	Data	Drivers	Sites	Parameters	States
1	FLUXNET NEE and LAI data, 2000–2001	Observed, 2000–2001	DE-EC, EV-EC	Generated by MDF, with 90% CI	Generated by MDF with 90% CI
2	Artificial/synthetic	Artificial	DE-SYN, EV-SYN	Generated by MDF, with 90% CI	Generated by MDF with 90% CI
3	None	Observed, 2002	DE-EC, EV-EC	From experiment 1	Generated by MDF with 90% CI
4	None	Artificial	DE-SYN, EV-SYN	From experiment 2	Generated by MDF with 90% CI

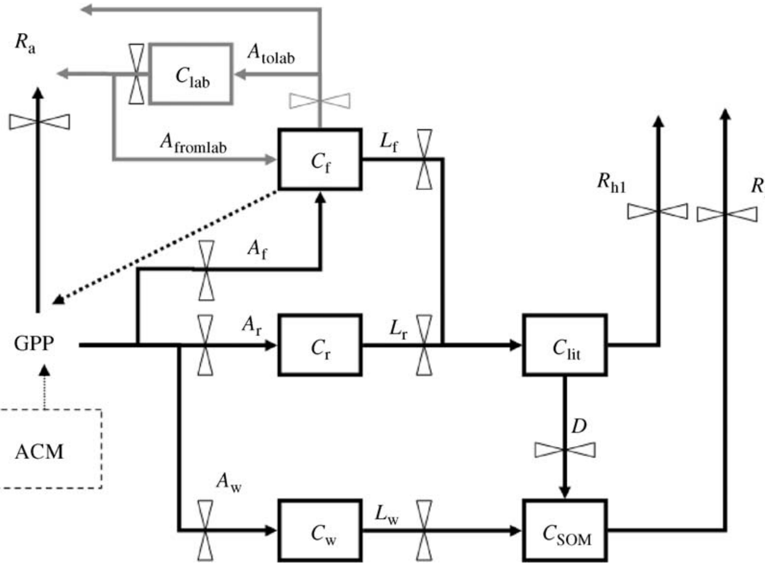


Fig. 1. A schematic of the DALEC (black) and DALEC-deciduous (black and grey) models. The figures show pools (boxes) and fluxes (arrows) of C. Feed-back between DALEC and the model of gross primary production, ACM, is indicated by dotted line. Allocation fluxes are A, litter-fall fluxes are L, and respiration is R, split between autotrophic (a) and heterotrophic (h). D is decomposition and GPP is gross primary productivity. C stocks: C_{lab} = labile; C_f = foliage; C_r = fine roots; C_w = wood; C_{lit} = litter; C_{SOM} = soil organic matter. Allocation: A_{tolab} = to labile pool; $A_{fromlab}$ = from labile pool, A_f = to foliage; A_r = to fine roots; A_w = to wood. Litterfall: L_f = from foliage; L_r = from fine roots; L_w = from wood.

4. For the deciduous model, the timing of initial leaf out is controlled by a simple growing degree day accumulation, and leaf fall by a minimum temperature threshold. The maximum amount of C that can be allocated to leaves is also limited by a parameter (C_{fmax})
5. All C losses are via mineralisation (i.e. no dissolved losses).

The aggregated canopy model (ACM) (Williams et al., 1997) is used to calculate daily GPP in DALEC. ACM is a 'big leaf', daily time-step model that estimates GPP using a simple aggregated set of equations operating on cumulative or average values of leaf area index (LAI, determined directly from C_f, total foliar C), foliar nitrogen, total daily irradiance, minimum and maximum daily temperature, day length, atmospheric CO_2 concentration, water potential gradient (ψ_d) and total soil–plant hydraulic resistance (r_{tot}). ACM contains 10 parameters (Table 4) which have been calibrated using a fine-scale model (the soil–plant–atmosphere model (SPA), (Williams et al., 1996) across a wide range of driving variables producing a 'universal' parameter set which maintains the essential behaviour of the fine-scale model but at a much reduced complexity.

2.2. Selection of parameters for optimisation

Across most carbon dynamics models (Sitch et al., 2008) a similar photosynthesis scheme is applied (Farquhar and von Caemmerer, 1982), the same as that used in SPA. There is strong consensus that this photosynthesis model is appropriate given current knowledge, and for this reason the REFLEX study optimised only one of the photosynthesis parameters used in DALEC. Rather, REFLEX focussed on optimising the parameters that determine how photosynthate is allocated, turned over and mineralised. These processes, and thus their parameters, remain uncertain, and yet are key determinants of NEE. For instance, the suggestion that the ratio of net primary production to GPP is constant (Waring et al., 1998) remains contested (Zhang et al., 2009), so there is uncertainty in the parameterisation of autotrophic respiration, to which NEE is highly sensitive (Williams et al., 2005). Phenological models for timing of leaf out and leaf senescence remain highly empirical but critical determinants of C balance (Van Wijk et al., 2003). Existing models for the allocation of photosynthate to wood and fine roots, and turnover of soil organic matter and fine roots use varying assumptions and parameters (Davidson and Janssens, 2006; Dewar et al., 2009), and generate different NEE estimates as a result (Sitch et al., 2008). For these reasons, the optimisation is directed at parameters determining the fate of fixed carbon, including the variation in leaf area in time, which is itself a critical determinant of photosynthesis.

The sole ACM parameter included in the optimisation is the nitrogen use efficiency parameter (a_1), which determines the maximum rate of carboxylation per g foliar N. For the purposes of this experiment the sites were treated as being non-drought stressed. Those variables related to drought effects in ACM, specifically ψ_d and r_{tot}, were given a fixed value in accordance with this assumption. Although only one ACM parameter is directly optimised, LAI dynamics create a strong feedback between DALEC and ACM (and thus GPP), which a number of parameters influence through allocation and turnover rates.

2.3. Datasets

Four datasets (two synthetic and two based on actual measurements) were provided to participants. Each dataset included a variety of information (Table 2), including continuous daily meteorological drivers, intermittent NEE and LAI data. Estimates of the initial values of the pools of soil organic matter and wood, and site data on leaf mass per area, for calculating LAI in the GPP model, were provided (Table 3). Initial conditions for foliar, fine root, litter and labile C were not provided, nor were expected ranges. No other information was provided.

Table 2
Time series data available provided to users in the experiments for all sites.

Observation	Units	Interval	Source
Global radiation	MJ m^{-2} day^{-1}	Daily	FLUXNET data portal
Minimum temperature	°C	Daily	FLUXNET data portal
Maximum temperature	°C	Daily	FLUXNET data portal
Atmospheric CO_2 concentration	μmol mol^{-1}	Daily	FLUXNET data portal
NEE	gC m^{-2} day^{-1}	Daily	FLUXNET data portal
LAI	m^2 m^{-2}	When available	References/site PI

Table 3
Site details, including latitude, initial conditions for large C pools, and foliage parameters, all provided to users.

Site	Latitude (°N)	Soil organic matter C (gC m^{-2})	Above-ground biomass (gC m^{-2})	Leaf mass per area (gC m^{-2} leaf area)	Foliar N (g N m^{-2} leaf area)
EV-EC (Loobos)	52	11000	9200	110	4.0
EV-SYN	50	9700	12400	110	3.8
DE-EC (Hesse)	48	7100	8800	22	1.0
DE-SYN	51	9900	8900	22	1.1

Synthetic datasets were generated for three years for an evergreen (EV-SYN) and deciduous (DE-SYN) forest, using DALEC and DALEC-D model runs, with nominal parameters and meteorological driver data selected from European EC flux tower sites (Viesalm, Belgium and Braaschaat, Belgium). Gaps were introduced into the synthetic NEE and LAI data time series by thinning the model outputs to match the data availability from the real data discussed below. The synthetic NEE data were first thinned to match the days for which daily NEE could potentially be determined when using a strict criteria of >43 of a possible 48 half-hour NEE observations passing quality control for the FLUXNET sites which were used to provide meteorological drivers (~40% of total time period). This coverage was found to be greater than that determined for the other FLUXNET sites used in the real data experiments (discussed below), and so to aid comparison these data were further thinned by removing days at random to reduce coverage to ~30%. Similarly, synthetic LAI data were only provided on days which mirrored the patterns of actual data availability at the real data experiments.

Noise was added to the remaining model outputs to reflect measurement error, by adding Gaussian errors with a variance of 0.5 g C m^{-2} day^{-1} for the NEE and 10% of the truth for the LAI (participants were not provided with these details). Though the half-hourly measurements may have non-Gaussian errors, the noise on the sum/mean becomes Gaussian with aggregation at longer time scales. Participants were provided with the first two years of synthetic observations.

For the observed data, the sites (Loobos, Netherlands and Hesse, France) and site-years (2000–2002) were selected on the basis of relatively long, continuous records of fluxes and site meteorology, quality controlled data, and little or no drought stress. The observed data included measurements of eddy covariance (EC) fluxes, LAI, and local meteorology from a deciduous broad-leaf forest (identified as DE-EC = Hesse) and an evergreen needle-leaf forest (EV-EC = Loobos). Daily NEE was calculated by summing half-hourly observations, but only if >43 of the possible 48 observations passed quality control. Missing data were filled by the daily mean of remaining data. It is possible that some small bias was introduced by this simple gap-filling, but for the purposes of this study such impacts were deemed insignificant. Typical data coverage was 20–30% of days. LAI data were sparse, usually collected on just a few days. Gap-filled flux data were not used in this experiment, but complete daily meteorological data were required to drive the model, and so gap-filled weather data were used. All data were obtained via the FLUXNET site (www.fluxnet.ornl.gov), from relevant, site-specific literature and/or from site PIs. Three sequential years of data were assembled, of which the first two years were provided to participants. The source of the EC data, and any estimated uncertainties were withheld from participants.

2.4. Experiments

All participants used DALEC and DALEC-D, the same models used to generate the synthetic data. The use of common reference models allowed direct comparison among MDF algorithms. Upper and lower bounds for the parameters of both deciduous and evergreen versions of the model were provided (Table 4). These bounds were set broad to ensure a high likelihood that reasonable parameters were located in the EC experiments. The ranges for turnover rates were set to give mean residence times of at least 10 days for leaves, labile C and litter and 100 days for other pools, and at most 27 years for foliage, labile C and roots, 270 years for litter and 2700 years for SOM and wood. The fraction of GPP respiration autotrophically is often set at ~0.5 (Waring et al., 1998), but other values have been reported so we set a range from 0.2 to 0.7. For allocation to foliage and roots we allowed these parameters to vary from very small values to a maximum of 0.5 each, so that total allocation to roots and leaves ≤1. The temperature response parameter (E_t) bounds were set to keep Q10 between 1.65 and 7.4, spanning the commonly found values between 2 and 3 (Davidson and Janssens, 2006). The GPP scalar (P_r) was set to allow approximate doubling or halving of the expected value (~10). The range of the L_{out} and L_{fall} parameters were set to span the expected *f* months of leaf out in temperate climates (March–May) and for the start of leaf abscission (September–November). Maximum foliar C (C_{fmax}) was set within a range to give LAI of ~5–25 for broadleaves and ~1–5 for needles, with this variability resulting from the differences in leaf mass per area for the different plant functional types. For fraction of leaf loss transferred to litter (F_{ll}), we set a range that resulted in 20–70% of foliar C being stored in the labile pool to prime growth in the following season.

Participants applied the MDF algorithm of their choice to four experiments (Table 1). The first two experiments were diagnostic, testing parameter and state estimation using two years of incomplete daily NEE and LAI data, at both an evergreen and deciduous site. These data were either real, collected at a FLUXNET site (experiment 1) or artificial, synthesised from model output with added noise (experiment 2). The final two experiments were prognostic, testing forecast capability, again at the real sites (experiment 3) and the artificial sites (experiment 4). Forecasts of daily C fluxes and pool dynamics were generated using parameter distributions from the first two experiments, forced by a single extra year of meteorological data. The flux/stock data, both observed and synthetic, for this third year were withheld for later assessment.

2.5. Algorithms

A wide range of different MDF algorithms are currently applied (e.g. Raupach et al., 2005), but because REFLEX was an open inter-comparison experiment the algorithms employed were not selected according to any criteria, rather they were dependent upon the community interest and experience (Table 5). Many of the methods used Monte Carlo approaches based on the Metropolis–Hastings algorithm or variants thereof. There were differences in the implementation, with various cost functions, uncertainty specifications and convergence tests employed. The cost function weights the difference between observations and simulated quantities, often using observation error estimates, and sometimes model error estimates. There was also a genetic algorithm approach, and an Ensemble Kalman Filter (EnKF). In two cases a Metropolis approach was supplemented by a Kalman filter (one Unscented KF, one EnKF).

Table 4
Model parameters for DALEC. p1–17 and a1 require calibration. NPP_2 is NPP remaining after allocation to foliage.

	Description	Code	Nominal value or range (low/high)
p1	Decomposition rate (per day)	T_d	1×10^{-6}/0.01
p2	Fraction of GPP respired autotrophically	F_g	0.2/0.7
p3	Fraction of NPP allocated to foliage	F_{nf}	0.01/0.5
p4	Fraction of NPP_2 allocated to roots	F_{nrr}	0.01/0.5
p5	Turnover rate of foliage (per day)	T_f	1×10^{-4}/0.1
p6	Turnover rate of wood (per day)	T_w	1×10^{-6}/0.01
p7	Turnover rate of roots (per day)	T_r	1×10^{-4}/0.01
p8	Mineralisation rate of litter (per day)	T_l	1×10^{-5}/0.1
p9	Mineralisation rate of SOM/CWD (per day)	T_s	1×10^{-6}/0.01
p10	Parameter in exponential term of temperature dependent rate parameter	E_t	0.05/0.2
p11	Nitrogen use efficiency parameter (a1) in ACM	P_r	5/20
p12*	*GDD* value causing leaf out	L_{out}	200/400
p13*	Minimum daily temperature causing leaf fall	L_{fall}	8/15
p14*	Fraction of C in leaf loss transferred to litter	F_{ll}	0.2/0.7
p15*	Turnover rate of labile carbon (per day)	T_{lab}	1×10^{-4}/0.1
p16*	Fraction of labile transfers respired	F_{lr}	0.01/0.5
p17*	Maximum C_f value (gC m^{-2})	C_{fmax}	100/500
a1	Nitrogen use efficiency of GPP		5/20
a2	Daylength coefficient		0.0156
a3	Canopy CO_2 compensation point		4.22
a4	Canopy CO_2 half saturation point		208.9
a5	Daylength constant		0.0453
a6	Hydraulic coefficient		0.378
a7	Maximum canopy quantum yield		7.19
a8	Temperature coefficient		0.011
a9	LAI-canopy quantum yield coefficient		2.10
a10	Water potential constant		0.79

* Parameters p12–17 are used in DALEC-deciduous only. a1–10 are parameters for the GPP model ACM used in DALEC.

All the algorithms (bar the free-standing EnKF) used ~10^5 iterations to produce the full set of parameter and state estimates. Most of the algorithms assumed that prior parameter distributions were uniform across the range supplied. The use of a uniform prior suggests that the researcher has a prior belief that all setting of parameters within the range are equally likely. The users made a variety of assumptions about initial conditions for some state variables (Table 5) and more detailed are provided in the Appendix A.

For the Metropolis methods, confidence intervals on fluxes were generated as a function of the set of acceptable parameter sets. These parameters sets were fed into the model to produce a set of possible outcomes, that were then sampled to determine the 90% CI. Differences in the size of the CI depend on the accept/reject criterion employed by each algorithm in generating acceptable parameter sets (Table 5). The methods employing the Kalman filter employed a further step, once acceptable parameter sets were determined. The state variables of the model, including flux estimates, were updated using sequential assimilation of observations through the times series.

2.6. Analyses

Because of the multiple datasets and algorithms employed, a series of metrics were required to most simply describe the outcomes of the parameter estimation exercises. To quantify and summarise the different approaches, we computed for each parameter two (for EC data) or three (for SYN data) relative-distance metrics, d_1–d_3. Here, for a given parameter, m_x is algorithm x's best estimate of the parameter; CI_x is the width of the parameter's confidence interval for algorithm x; t is the true value of the parameter; p_{max} and p_{min} are the pre-specified upper and lower limits on the parameter (Table 4); σ is a standard deviation and μ is a mean:

$$d_1.\quad \text{Consistency among algorithms}$$
$$:\quad \sigma(m_1, \ldots, m_9)/(p_{max} - p_{min})$$

This metric tests whether all algorithms retrieve a similar signal from the observations. It does not indicate whether the retrieved parameter is "correct", but quantifies the ability of algorithms to find a consistent part of the parameter space and identify minima.

$$d_2.\quad \text{CI constrained by the data}: \quad \mu(CI_1, \ldots, CI_9)/(p_{max} - p_{min})$$

This metric tests to what degree the posterior estimate of the parameter is an improvement on the prior estimate. The size of d_2 will depend on the prior estimate range, and so this metric has a subjective component.

$$d_3.\quad \text{Consistent with truth (SYN only)}$$
$$:\quad |t - \mu(m_1, \ldots, m_9)|/(p_{max} - p_{min})$$

This metric actually determines whether the retrieved parameter is consistent with the truth, which is known for the SYN case.

We determined two further metrics to aid a comparison among algorithms of parameter estimation capabilities, for the SYN cases only. Mean normalised parameter confidence interval (d_4) is similar to the d_2 statistic but rates individual algorithm's mean 90% confidence intervals across all parameters, normalised by the size of the parameter priors:

$$d_4.\quad \left(\sum_{i=1}^{n} \frac{CI_i}{p_{max\,i} - p_{min\,i}}\right)/n$$

where CI_i is the width of the algorithm's 90% confidence interval for parameter x; n is the number of parameters (11 for EV, 17 for DE), $p_{max\,i}$ and $p_{min\,i}$ are the pre-specified upper and lower limits on each parameter prior.

The metric for consistency with true parameter value (d_5) is similar to the d_3 statistic, but again rates consistency for an individual algorithm across all parameters:

$$d_5.\quad t - \left(\sum_{i=1}^{n} \frac{m_i}{p_{max\,i} - p_{min\,i}}\right)/n$$

where m_x is parameter x's best estimate by the algorithm and n is the number of parameters. The closer d_5 is to zero, the better.

Table 5
A summary of the algorithms used in the experiment. Methods using Metropolis algorithm alone are labelled Mx. U1 and E1 used a Kalman filter after an initial Metropolis algorithm search for parameters. G1 and E2 are the only methods not using the Metropolis algorithm in the some manner. G1 did not generate confidence intervals for GPP and R_e.

Participant	Name – type of methodology	Code	Prior	Cost/objective function	Initial pools	Convergence tests	Number of parameter sets produced	Number of model iterations	Programming language
E1 (stage 1)			Uniform	Weighted root mean square error	Parameters to be estimated	Gelman and Rubin (1992)	~400000	~1000000	Fortran
E1 (stage 2)	MCMC Metropolis, then EnKF	Evensen (2003)	PDFs from stage 1	Kalman gain	PDFs from stage 1	n/a	State only	8000	Fortran
E2	Ensemble Kalman Filter	Evensen (2003)	Gaussian	Kalman gain	Included in calibration	n/a	~2000	800	Fortran
U1	Unscented Kalman Filter	Gove and Hollinger (2006)	Gaussian	Minimize posterior error covariance via the Kalman gain.	As estimated by M3	n/a	State only	n/a	R
G1	Genetic algorithm	Based on Haupt and Haupt (2004)	Uniform	Based on Haupt and Haupt (2004)	Tuned with parameters	n/a	~100000		Fortran
M1	MCMC – Metropolis			Gaussian likelihood	Included in calibration	visual		300000	Fortran
M2	MCMC – Metropolis	MCMC1	Uniform	Weighted root mean square error	Parameters to be estimated	Visual comparison of parameter PDFs from 2 chains	1000000	1000000	Fortran
M3	Simulated annealing – Metropolis	SAM	Uniform	Weighted root mean square error	Parameters to be estimated	n/a	1000	~250000	Fortran
M4	MCMC – Metropolis	MCMC3	Uniform	Weighted root mean square error	Spinup to equilibrium of total C	Heidelberger and Welch (1983)	80000	~300000	R
M5	Multiple complex MCMC – Metropolis	SCEM	Uniform	Weighted root mean square error	Parameters to be estimated	Gelman and Rubin (1992)	~500000	150000	Matlab

3. Results

3.1. Parameter estimation

Each algorithm produced sets of parameter estimates for each dataset in experiments 1 and 2, describing a multi-dimensional probability density volume. Because of their high dimensionality, these hyper-volumes are not easily described or visualised, so a range of metrics and methods are used. Firstly, we determined the "best" parameter set estimate of each algorithm (Fig. 2), based on the minimum of the cost function (e.g. Metropolis algorithm) or the mean value of an ensemble (Ensemble Kalman Filter). The best estimates were supplemented by estimates of the 90% confidence intervals (CIs) on each parameter. These CIs were calculated on the basis of the 5.0 and 95.0 percentiles of the accepted parameter distributions submitted by participants.

For the SYN datasets only, it was possible to gauge how effectively the algorithms retrieved the true parameter values using d_3 (Table 4). The analysis reveals (Table 6, Fig. 2 and Appendix A) that turnover rate parameters for soil (T_s), foliage (T_f) and litter (T_l) as well as the temperature rate parameter (E_t) and the NPP:GPP ratio (F_g) were well estimated overall, across the range of methods for deciduous and evergreen ecosystems. By comparison, the turnover rate parameters for wood (T_w) and decomposition (T_d), tended to be poorly estimated overall. The allocation to foliage parameter (F_{nf}) was well estimated for EV-SYN but biased in DE-SYN. Of those parameters used only in the deciduous model, allocation to litter (F_{ll}) and turnover of labile C (T_{lab}) were poorly estimated, whereas the labile transfer respiration fraction (F_{lr}) was more successfully estimated. The estimates of the phenology parameters L_{out} and L_{fall} were of intermediate quality.

For the SYN datasets, the d_2 metric indicated that several of the turnover rate parameters were well constrained compared to their priors (i.e. T_f, T_l and T_s) with narrow confidence intervals (Table 6). The d_1 metric indicates that these same parameters were consistently estimated among algorithms. Conversely, some parameters were poorly constrained, e.g. F_{nf}, with little reduction in spread from the initial upper and lower bounds (d_2 metric), although in this case, as already noted, the best estimate values were close to the truth for the EV-SYN case. The d_1 metric revealed that several parameter estimates were not consistent among algorithms, including maximum leaf area parameter (C_{fmax}) in DE-SYN.

Table 6
Parameter estimation metrics using nine different algorithms based on synthetic data for evergreen (left) and deciduous (right) forest. Metric d_1 quantifies consistency among methods; d_2 quantifies the data constraint on the confidence intervals; and d_3 quantifies the consistency with the truth.

	Evergreen: EV-SYN			Deciduous: DE-SYN		
Param	d_1	d_2	d_3	d_1	d_2	d_3
T_d	0.26	0.36	0.75	0.26	0.42	0.72
F_g	0.30	0.41	0.02	0.11	0.42	0.09
F_{nf}	0.07	0.49	0.00	0.26	0.53	0.37
F_{nrr}	0.24	0.65	0.31	0.19	0.60	0.07
T_f	0.06	0.20	0.03	0.05	0.16	0.01
T_w	0.22	0.40	0.69	0.27	0.37	0.22
T_r	0.27	0.52	0.03	0.04	0.28	0.02
T_l	0.07	0.22	0.03	0.03	0.15	0.03
T_s	0.05	0.16	0.21	0.04	0.08	0.01
E_t	0.04	0.24	0.00	0.05	0.17	0.04
P_r	0.21	0.47	0.15	0.14	0.46	0.06
L_{out}				0.22	0.40	0.19
L_{fall}				0.14	0.25	0.10
F_{ll}				0.13	0.52	0.24
T_{lab}				0.19	0.54	0.01
F_{lr}				0.18	0.33	0.00
C_{fmax}				0.22	0.36	0.17
Mean	0.16	0.38	0.20	0.15	0.36	0.14

A critical parameter controlling C accumulation is F_g (Williams et al., 2005), which determines what fraction of GPP is respired by plants. While in both SYN cases the d_3 metrics indicated reasonable consistency with the truth, the d_2 metrics indicated relatively poor constraint on parameter confidence intervals by the data and the d_1 metrics indicated a degree of inconsistency among algorithms. The d_1 and d_2 metrics were similarly above average for the allocation parameters to foliage (F_{nf}) and roots (F_{nrr}), indicating relatively poor agreement among algorithms and limited constraint on priors.

For the EC datasets the 'true' parameter value is not known so it is only possible to comment on the levels of parameter constraint and consistency between algorithms (Table 7 and Appendix A) and make comparison with the SYN cases. Again the d_2 metric indicates that most of the turnover rate parameters (i.e. T_f, T_l and T_s) and the temperature response of respiration (E_t) were well constrained. But in general there was more variability between algorithms in the CIs than in the SYN cases. Consistency between algorithms, although in no way a measure of how good the algorithms are at identifying the

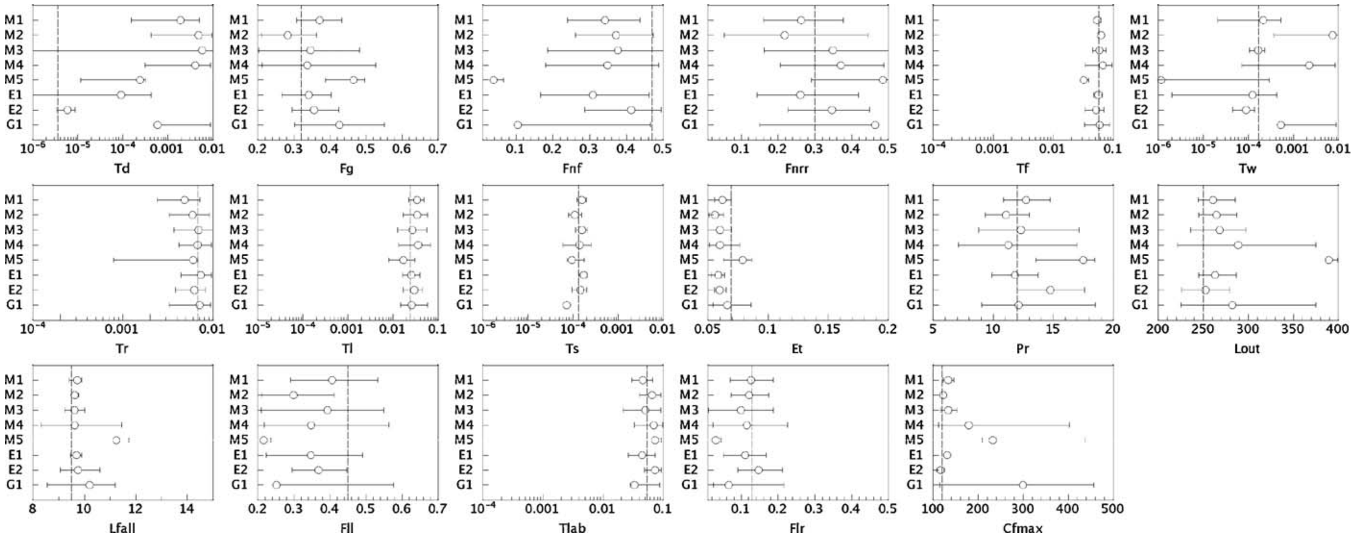


Fig. 2. Parameter estimation for deciduous synthetic (DE-SYN) data. The panels shows each of the algorithms' best estimate of each parameter, and the magnitude of each 90% confidence intervals. The 'true' value of the parameter used in generating the synthetic data is indicated by the *d* vertical line. The upper and lower bounds of each parameter, as provided to the experimenters, is indicated by the range of each *x*-axis. *x*-Axes are log scaled for turnover rates (all parameters beginning *T*). For an explanation of parameter symbols see Table 5.

Table 7
Parameter estimation metrics using nine different algorithms based on observed data for evergreen (left) and deciduous (right) forest. Metric d_1 quantifies consistency among methods; d_2 quantifies the data constraint on the confidence intervals.

Evergreen EV-EC			Deciduous DE-EC	
	d_1	d_2	d_1	d_2
T_d	0.28	0.42	0.29	0.36
F_g	0.11	0.36	0.08	0.3
F_{nf}	0.16	0.31	0.2	0.55
F_{nrr}	0.29	0.6	0.15	0.53
T_f	0.08	0.19	0.12	0.25
T_w	0.24	0.35	0.21	0.35
T_r	0.29	0.35	0.32	0.2
T_l	0.09	0.23	0.08	0.18
T_s	0.08	0.1	0.05	0.2
E_t	0.02	0.2	0.09	0.19
P_r	0.14	0.52	0.17	0.35
L_{out}			0.21	0.37
L_{fall}			0.2	0.32
F_{ll}			0.16	0.32
T_{lab}			0.1	0.49
F_{lr}			0.12	0.23
C_{fmax}			0.03	0.25
Mean	0.16	0.33	0.15	0.32

correct values, is an indication of which parameters the algorithms find most identifiable given the observations and model structure. For EV-EC greatest consistency is shown for E_t whilst for DE-EC it is the deciduous model only parameter, maximum foliar C, C_{fmax}. For a number of parameters (i.e. T_d and L_{fall}) the algorithms seem to split their best estimate values into two groupings. This result indicates different minima were found by the different algorithms and potential problems with equifinality through parameter covariance.

Comparing the SYN and EV results for parameter consistency there was a significant correlation in the associated metric (d_1) between EC and SYN for EV ($r = 0.73$, $P = 0.01$) but not DE ($r = 0.31$, $P = 0.24$). So the EV parameters that were consistently estimated (across methods) were similar for synthetic and eddy covariance data, while this was not so for DE datasets, perhaps because of the greater number of parameters. There was a significant correlation between EC and SYN d_2 distances, measuring how well parameter CIs were constrained by data, for both EV ($r = 0.87$, $P = 0.0004$) and DE ($r = 0.84$, $P < 0.0001$). Thus parameters that were well constrained (low d_2) by the synthetic data were well constrained by the eddy covariance data.

As expected, those algorithms with large parameter confidence intervals encompassed a large fraction of true parameter values within their 90% confidence intervals for the SYN cases (Fig. 3). For the DE-SYN case, three algorithms (E1, E2, M1) managed to generate relatively small and reliable confidence intervals. For the EV case, none of the algorithms managed to balance small confidence intervals with reliability. For the DE case, three algorithms (E1, E2, M1) generated parameters that were most consistent with true values and also had the smallest confidence intervals. For the EV case there was no clear pattern among algorithms; although E2 had the closest agreement with true parameters and the narrowest confidence intervals, it had the smallest fraction of true parameters within the 90% CI, suggesting over-confidence.

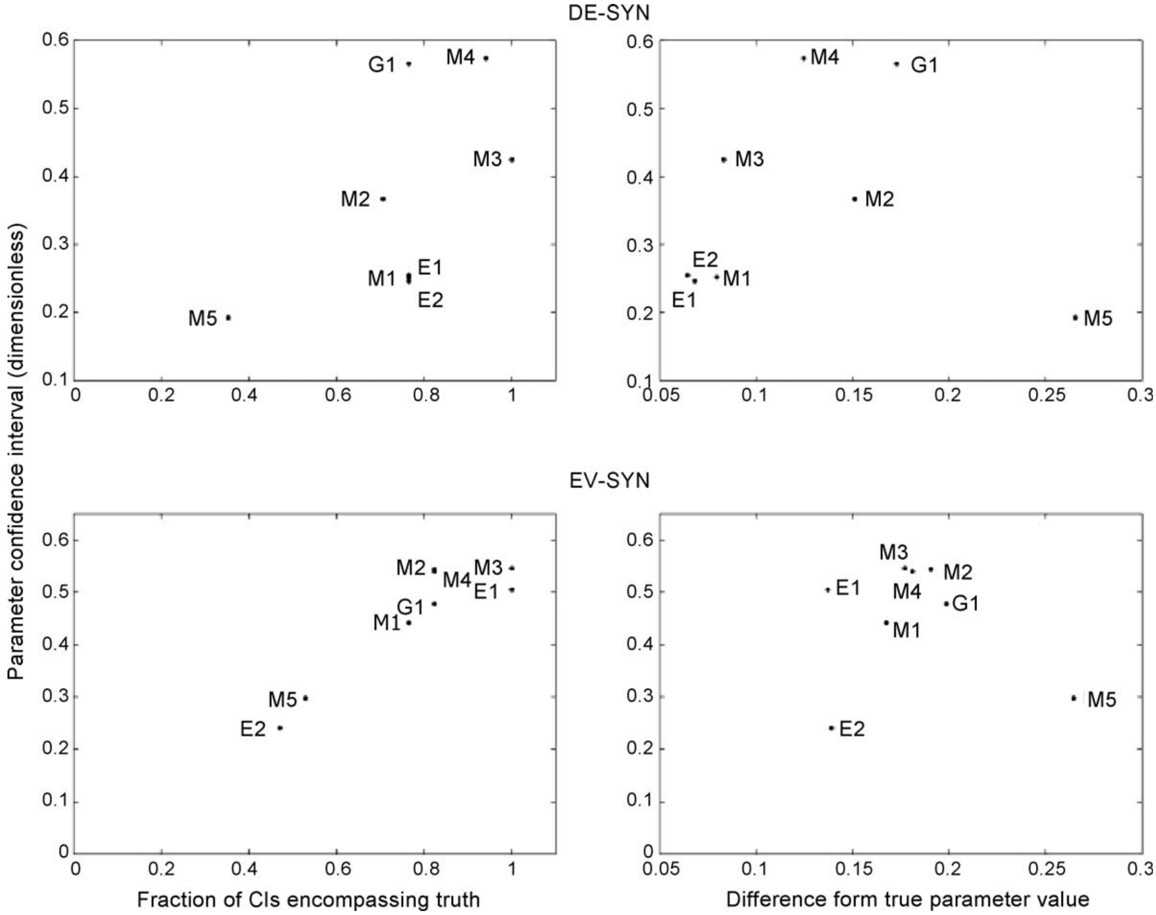


Fig. 3. A comparison of two metrics of parameter calibration success against mean parameter 90% confidence intervals of each algorithm (d_4, see text). Parameter calibration success is judged in two ways: (1) by the fraction of 90% confidence intervals encompassing the true parameter values obtained by each algorithm, see left panels—high values are better; (2) by the mean normalised difference between best estimate and true parameter values obtained by each algorithm (*d5*, see text), see right panels—low values are better. Individual algorithms are identified by alphanumerics (Table 5). The top two panels are generated from the deciduous synthetic data, the bottom two from evergreen synthetic data. Data for the synthetic experiments are shown, where true values of the parameters are known.

3.2. Flux estimates—synthetic data

For the synthetic datasets daily NEE predictions were generally close to the true values from which observations were generated. RMSE values calculated against the true values ranged from 0.07 to 0.55 gC m^{-2} day^{-1}, with a mean over all algorithms and years of 0.20 gC m^{-2} day^{-1}. These error values compared well with the noise added to the truth in order to generate synthetic observations and can be compared with mean NEE true value of -0.44 gC m^{-2} day^{-1}.for EV-SYN and 0.01 gC m^{-2} day^{-1}.for DE-SYN. This would suggest that for the EV-SYN case the algorithms are correctly able to identify the site as a carbon sink, but for the DE-SYN site which is very near equilibrium they would necessarily be able to attribute the small C source correctly. Partitioning synthetic NEE into GPP and R_e was generally successful compared to the known true values, with mean RMSE values over all algorithms of 0.6 gC m^{-2} day^{-1} in both cases, which can be compared with GPP and R_e of 3.4 and 3.41 gC m^{-2} day^{-1} for DE-SYN and 2.26 and 1.82 gC m^{-2} day^{-1} for EV-SYN, respectively. There was no evidence that best-fit or mean predictions of fluxes deteriorated in year 3, the prognostic period during which data were not assimilated.

3.3. Parameter correlations

Model structure can generate correlations among parameters, so in some cases the same model output can be generated using different parameter values, i.e. equifinality (Richardson and Hollinger, 2005; Schulz et al., 2001). So, we determined whether the parameter estimates of the different algorithms were similarly correlated. Parameter correlations indicate a potential weakness in constraining the parameters concerned, and we expected that correlated parameters would have broader confidence intervals as a consequence. We compared the parameter correlation matrices produced by each algorithm for each of the four datasets. From each matrix we found the 5 highest absolute values, to simplify the analysis. We then determined how many of these top five correlations were in common among algorithms ($n = 9$) for each dataset. The number of unique ranked correlations could vary from 5 (all algorithms in agreement) to 45 (no agreement). The observed numbers varied from 21 to 24.

For the EV datasets, 8 out of 9 algorithms agreed on a high ranking for a correlation between allocation to foliage and turnover rate of foliage (F_{nf} and T_f). Also for the EV datasets, five algorithms ranked highly a correlation between the fraction of GPP respired and the photosynthetic rate parameter (F_g and P_r). 8 out of 9 algorithms rated this correlation highly for DE-SYN, but only two algorithms for DE-EC. For both DE-EC and DE-SYN the algorithms agreed on an important correlation between allocation to fine roots and the turnover rate of SOM (F_{nrr} and T_s, identified by five algorithms) and between turnover rate of foliage and trigger for leaf fall (T_f and L_{fall}, identified by four algorithms).

An eigenvector analysis of the parameter covariance matrix suggested that the best constrained parameter was the turnover rate of SOM, T_s. The next best constrained parameter identified was the temperature rate parameter, E_t. Turnover rate of foliage was well constrained for EV analyses. Allocation to and turnover of roots were poorly constrained for EV analyses. The results for the DE analyses were less clear, with differences between DE-EC and DE-SYN. Turnover rate of wood and roots were least well constrained in DE-SYN, while the GDD threshold for leaf out and the turnover rate of labile C were least well constrained in DE-EC. There was some variation in the eigenvectors from the different algorithms with some parameters well constrained by some algorithms, but not well constrained by others. Comparison with the constraint metric d_2 were largely, but not totally, consistent. Eigenvector analysis did not identify any consistent correlation features, apart from one between fraction of GPP respired (F_g) and the NUE parameter, P_r, consistent with earlier analyses.

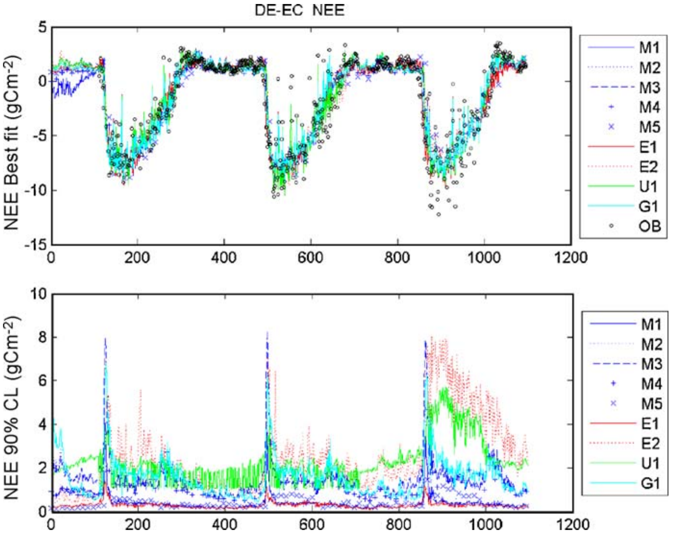


Fig. 4. Estimated time series of net ecosystem exchange of CO$_2$ (NEE) over three years from each algorithm using observations from the DE-EC dataset over the first two years (top panel) and 90% confidence intervals on the estimates (lower panel). The eddy covariance data is shown as open symbols.

3.4. Flux confidence intervals—daily data

The seasonal patterns of variation in NEE were generally well reproduced by most algorithms across all three years of each of the different datasets (for example, Fig. 4). There was low agreement among algorithms in the assessment of 90% confidence intervals (CI) on daily fluxes (Fig. 4). There were differences in confidence interval estimates both in magnitudes and in temporal variability among algorithms. For instance, the mean daily 90% CI varied among algorithms from 0.35 to 1.92 gC m^{-2} day^{-1} in DE-SYN and 0.29 to 2.49 gC m^{-2} day^{-1} in DE-EC. Algorithm confidence intervals typically had large excursions during spring leaf-out for DE, but the magnitude of these excursions varied (Fig. 4).

We tested whether the 90% CI on daily analyses (years 1 and 2) and predictions (year 3) encompassed the truth from the synthetic datasets for NEE, GPP and R_e for all years, and for observed NEE in year 3 for the EC datasets. The days of each year which passed this test were counted. We expected that 85–95% of the days would pass, roughly consistent with the magnitude of the confidence interval, 90%. For the synthetic experiments (NEE tests are shown in Table 8) this was rarely the case. In some cases the fraction was 100%, which indicates that the daily CI were likely set too large. In other cases, the fractions were <85% suggesting that the CI were too small or the predictions were biased. For the eddy covariance datasets in year 3, the majority of algorithms' confidence intervals on daily NEE were too narrow, with an average of only 40% (DE) or 20% (EV) of the observed year 3 data lying within the 90% confidence interval (Table 9). This result suggests the algorithm generated over-confident assessments of daily fluxes.

3.5. Flux estimates—observed data

For the eddy covariance datasets, the algorithms' predictions were compared to observed NEE. In years 1 and 2, when observations were provided to participants, RMSEs varied from 0.7 to 1.8 gC m^{-2} day^{-1} (DE) or 0.6 to 0.9 gC m^{-2} day^{-1} (EV), with a mean value of 1.3 gC m^{-2} day^{-1} for DE datasets and 0.7 gC m^{-2} day^{-1} for EV. In year 3, when observations were not provided to participants, RMSEs varied from 1.1 to 2.3 gC m^{-2} day^{-1} (DE) or 1.3 to

Table 8
Fraction of days in each year where 90% confidence interval encompassed the synthetic "true" value of NEE. Fractions are shown for each of the three individual years for DE-SYN and EV-SYN datasets. Values between 0.85 and 0.95 are in bold and are consistent with the 90% CI. Values of 1.0 are indicated by italics.

	DE-Syn			EV-Syn		
	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
Algorithm						
M1	**0.95**	0.97	0.99	0.81	**0.89**	1.00
M2	0.73	0.65	0.81	**0.95**	0.61	0.51
M3	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*	*1.00*
M4	**0.95**	0.97	0.96	0.80	**0.86**	**0.85**
M5	0.66	0.37	0.36	0.39	0.25	0.35
E1	**0.90**	0.83	**0.95**	0.93	0.77	0.69
E2	0.85	0.99	*1.00*	0.44	0.61	0.60
U1	0.99	0.99	*1.00*	0.99	0.98	*1.00*
G1	*1.00*	*1.00*	0.98	*1.00*	0.99	*1.00*

1.7 gC m^{-2} day^{-1} (EV), with a mean value of 1.5 gC m^{-2} day^{-1} for both EC and DE datasets (Table 9). Thus the best NEE estimates of the algorithms tended to agree less well in the prognostic period (year 3) compared to the assimilation period (years 1 and 2), though this was most striking for the evergreen (EV) case in this study.

3.6. Flux confidence intervals—annual sums

A comparison of 90% confidence intervals on annual estimates of NEE, GPP and R_e for all years revealed differences of up to an order of magnitude in size of CI (Fig. 6). There was no clear relationship between size of CI and algorithm type – for instance, M1 and M2 tended to have small CI compared to M3 and M4, although all used Metropolis algorithms. This result makes clear the importance of the user in determining the confidence interval, rather than the algorithm itself. The mean confidence interval for NEE (124 gC m^{-2} year^{-1}) was ~3-fold smaller than those for GPP (389 gC m^{-2} year^{-1}) and R_e (387 gC m^{-2} year^{-1}). A comparison of the mean 90% confidence intervals on annual NEE estimates (Table 10) indicated that CI were largest during year 3, the prediction period, and smallest in year 2. Of the 36 cases, 4 datasets, 9 algorithms, 34 had larger confidence intervals on year 1 than year 2, and 35 had larger CI on year 3 than year 2, so this pattern was general across algorithms and datasets. Averaged over all cases, the 90% CI in the prediction period (year 3) were 88% larger than in the second year of the assimilation period (year 2). Patterns were similar in comparison between outputs from observed and synthetic datasets. However, mean 90% CI across all algorithms were ~31% larger for EC datasets than for SYN datasets. Among algorithms, the increase in 90% CI on EC datasets compared to SYN datasets ranged from 0% (E1) to 100% (E2).

Table 9
Assessment of year three best-fit predictions and 90% confidence intervals (CI) for the EC datasets. Comparisons with both foliar carbon mass (C_f) and daily net ecosystem exchange (NEE) are shown. Assessment of best-fit predictions is through root mean square error (RMSE) on observations for year 3 for deciduous (DE) and evergreen (EV) forests. Assessment of confidence intervals is through quantifying the fraction of days in year 3 where the 90% confidence interval encompassed the observed NEE. Values between 0.85-0.95 are in bold and are deemed consistent with the 90% CI. Values of 0 indicated no observed data were within the CI, while a fraction of 1 indicates all data were within the CI. Algorithms are identified by codes. n is number of observations in year 3, which were withheld from the experimental team.

Algorithm	Foliar C mass (C_f)				Daily NEE			
	RSME (gC m^{-2})		CI frac		RMSE (gC m^{-2} day^{-1})		CI frac	
	DE-EC	EV-EC	DE-EC	EV-EC	DE-EC	EV-EC	DE-EC	EV-EC
M1	12.3	29.9	1	0.5	1.42	1.50	0.14	0.14
M2	7.6	16.6	0	0.83	1.21	1.34	0.11	0.06
M3	6.9	19.4	1	0.94	1.35	1.42	0.56	0.16
M4	16.9	18.9	1	1	1.57	1.73	0.39	0.19
M5	10.6	17.8	0	0.17	2.25	1.37	0.2	0.16
E1	6.1	20.3	1	0.33	1.10	1.49	0.14	0.08
E2	30.2	37.8	1	1	1.70	1.45	0.86	0.16
U1	4.1	15.5	1	1	1.34	1.37	0.84	0.61
G1	4.2	22.6	1	0.83	1.24	1.54	0.43	0.16
n	1	18	1	18	218	171	218	171

3.7. Testing annual flux estimates and confidence intervals

Annual flux outputs estimated and forecast using the synthetic datasets were compared with the synthetic truth. Each algorithm's annual output of NEE, GPP and R_e was tested to determine whether the truth lay within the 90% CI for estimates. The fraction of tests that were successful was compared with the mean size of the 90% confidence interval for each specific algorithm (Fig. 5). As expected there was often a positive relationship between success rate and confidence interval size, but some algorithms managed to contain the truth within relatively narrow confidence intervals. In the comparison for annual NEE, four algorithms (E1, E2, M1, M3) produced analyses with >80% success rate and mean confidence intervals <110 gC m^{-2} year^{-1}. In the comparison against component fluxes (GPP and R_e), two algorithms (E2, M2) produced more balanced analyses, with relatively high success rates (>65%) and narrow confidence intervals (<300 gC m^{-2} year^{-1}). M3 was always 100% successful in containing the truth within its 90% confidence intervals, and this over-confidence was because associated CI were the largest of all algorithms for GPP and R_e. There were successful tests for prognoses in year 3 by several algorithms, indicating that predictions of C fluxes beyond the observational period were successful also (Table 8).

3.8. GPP and R_e estimates

The decomposition of observed NEE data into GPP and R_e revealed major differences among algorithms, with best estimates varying by up to 900 gC m^{-2} year^{-1} (Fig. 6 and Appendix A). However there were similar patterns among algorithms across years. For instance, M4 tended to estimate lower magnitudes of these fluxes than other algorithms. In most cases the algorithms ranked the GPP and R_e similarly across years at each site, but not always. For instance, M1 and M5 ranked R_e differently for DE-EC across years (see Appendix A). Flux analyses were compared with estimates from other gap-filling and GPP-R_e decomposition algorithms using data from the same sites (Desai et al., 2008). In some cases there was close agreement between estimates, for instance NEE at Loobos in 2000 (Fig. 6), but in other, such as Loobos in 2001, there was disagreement.

3.9. Stocks

The analyses and predictions of foliar C matched the seasonal cycles and magnitudes of the truth from the synthetic studies adequately (Fig. 7). Predictions of year 3 foliar C in the eddy

Table 10
Mean size of 90% confidence interval on annual NEE for three years. Assessments were made with outputs from the nine algorithms, and compared for different years and datasets. The outputs for the first two years were analyses, based on model-data fusion. The output for the final year was generated from model predictions using estimated parameters and meteorological forcing, and no data. Units are gC m^{-2} year^{-1}.

Dataset	Year 1	Year 2	Year 3
DE-EC	181.0	96.6	186.2
EV-EC	119.3	92.6	169.4
DE-SYN	139.3	83.8	148.9
EV-SYN	95.4	58.1	117.9

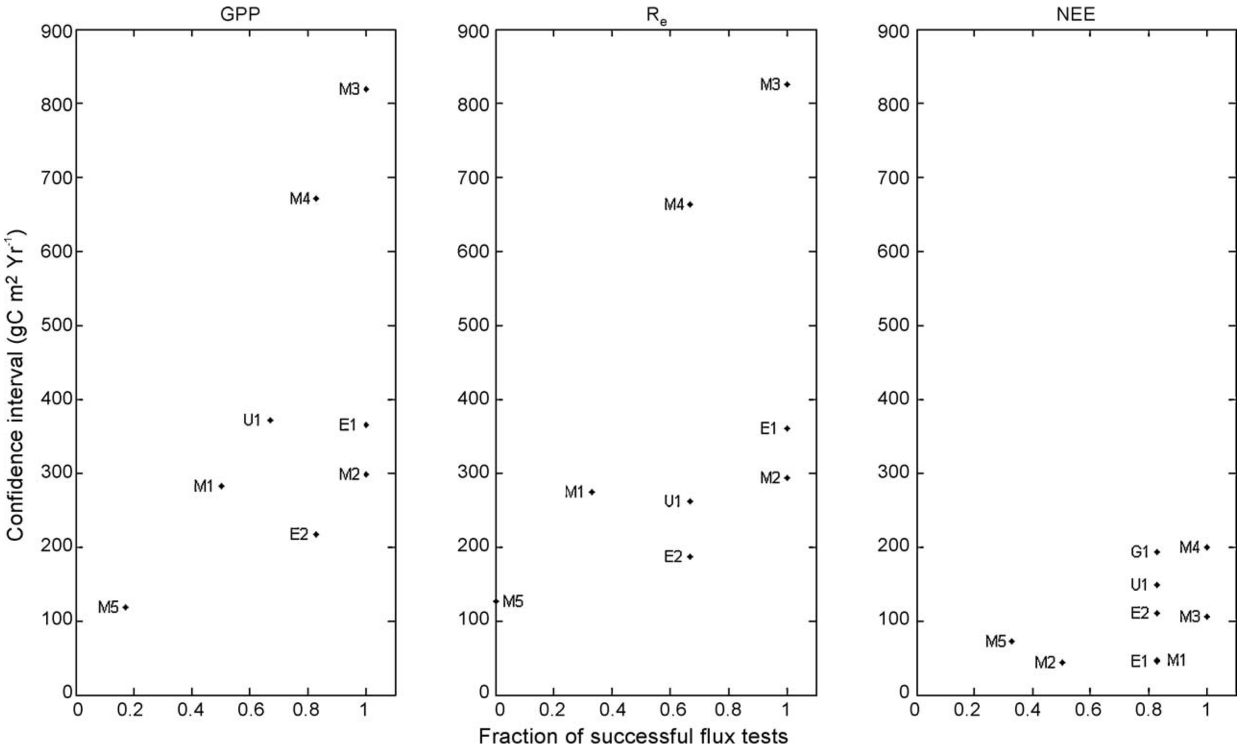


Fig. 5. A comparison between the summary success rate of annual estimates of GPP (left), R_e (centre) and NEE (right) for each algorithm plotted against the mean size of the 90% confidence interval used in the tests. The tests were for DE-SYN and EV-SYN, the synthetic datasets. Success was judged on whether each "true" annual flux was within the 90% confidence interval of the estimate. There were 6 tests (3 years × 2 datasets) for each flux. On the right panel the results for E1 and M1 were very similar. All panels have the same scale.

covariance datasets had a mean RMSE among algorithms of 11 gC m^{-2} for DE and 22 gC m^{-2} for EV. However, assessments of confidence intervals were generally poor; most algorithms had 90% CI either too broad or too narrow (Table 9).

For the synthetic data, the algorithms reproduced the seasonal cycles in fine root biomass, but the magnitude of the cycles and the mean biomass varied among the algorithms by ~50% (see Appendix A). This result reflected the choice of initial conditions,

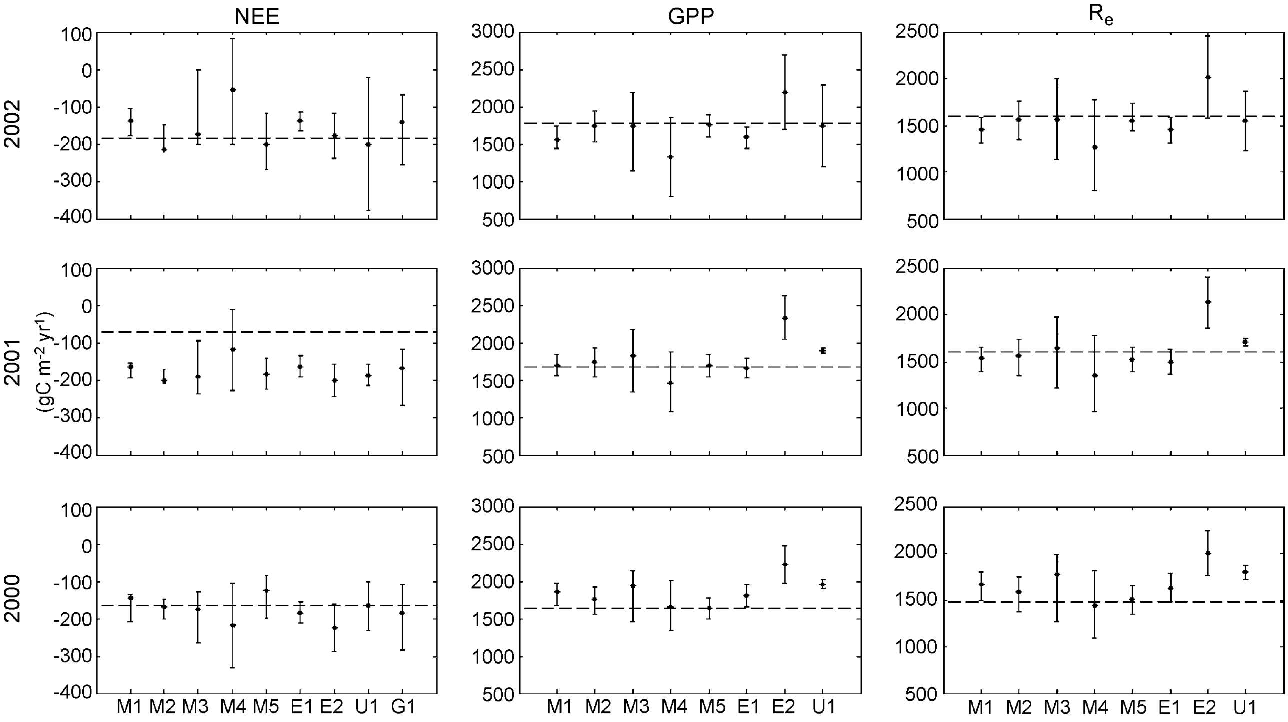


Fig. 6. Annual analyses of NEE, GPP and Re for 2000, 2001 and prognoses for 2002 generated with the EV-EC dataset from Loobos, Netherlands. Results are shown for each algorithm for NEE and for eight algorithms for GPP and R_e, with 90% confidence intervals indicated. The dashed lines show the best estimates from gap-filling routines using hourly NEE data (Desai et al., 2008).

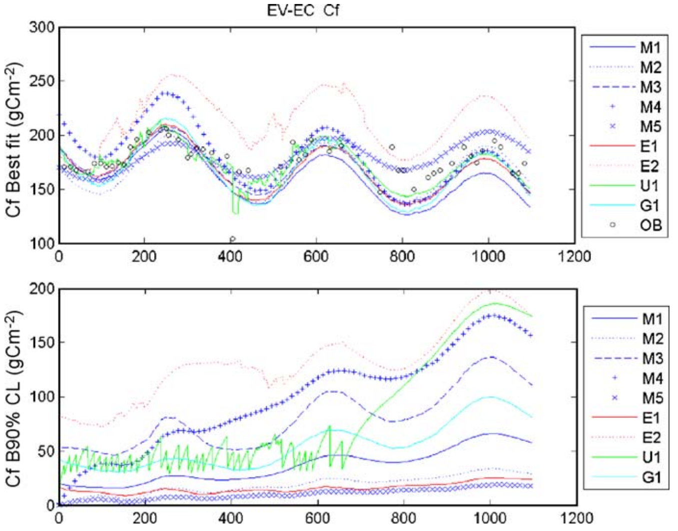


Fig. 7. Retrieved estimates of foliar C stocks over three years for the EV-EC deciduous site with observations of NEE fluxes and LAI assimilated. The upper panel shows best-fit or mean for C_f, with observations marked, and the lower panel shows the width of the 90% confidence interval.

or method of assessment, by the users. We found similar patterns in litter and labile C pools (see Appendix A).

There were some important differences in the analyses and predictions of the slow turnover C pools in all datasets. C_{som} in most analyses showed slight increases or decreases over time, but some algorithms showed stocks doubling over three years (M4 and M5, see Appendix A). Such doublings were unrealistic outcomes, but in these cases the algorithms were able to make these changes consistent with the flux observations. For C_w most algorithms suggest a small increase in C stocks over time, but the algorithms with increasing C_{som} (M4 and M5) matched this with decreases in C_w of similar magnitude.

4. Discussion

There have been previous attempts to parameterise C budget models using time series of C fluxes (Braswell et al., 2005; Knorr and Kattge, 2005; Wang et al., 2007). These studies have tended to focus on calibrating physiological parameters, related to photosynthetic and respiration rates, rather than parameters related to allocation and turnover of C pools. The calibration of parameters interacting on a range of timescales and links to data over several years is thus an important and novel component of REFLEX. The feedbacks between fluxes and stocks (e.g. photosynthesis and foliar C), and between soil organic matter and temperature, are particularly important determinants of NEE in the DALEC model that are investigated in REFLEX.

4.1. Parameter estimation and constraint

We expected that parameters linked to fast-response processes that mostly determine net ecosystem exchange of CO_2 (NEE) would be well estimated (d_3 metric) and well constrained (d_2 metric), while parameters for processes indirectly related to NEE would be poorly characterised. The parameters directly related to NEE are those related to GPP and autotrophic respiration (F_g, F_{nf}, T_f, P_r) and to turnover of litter and SOM (T_l, T_s, E_t). Parameters associated with dynamics of wood and fine roots are indirectly associated with NEE (i.e. only affect it through impacts on other vegetation pools). Phenology parameters will be important in the deciduous case, but largely at the spring and autumn shifts.

Our analyses largely supported our expectation. The NPP:GPP ratio (F_g), and the turnover of litter (T_l) and foliage (T_f) were well estimated in SYN cases according to d_3. These parameters are closely associated with foliage mass and/or gaseous exchanges of C, as they control magnitudes of autotrophic respiration and LAI dynamics. The turnover rate of SOM (T_s), a large slow turnover pool, was well estimated by most algorithms (Table 6, Table 7). Most algorithms were able to estimate the turnover of T_l and T_s and thus predict R_h from the NEE data, i.e. separate the two fluxes R_{h1} and R_{h2} (Fig. 1). We suggest this success was due to the relatively large seasonal variation in the litter pool size compared to SOM, and the provision of SOM initial condition data, allowing the two signals to be retrieved separately. Parameters associated with the turnover of wood (T_w) and allocation to roots (F_{nrr}) were poorly estimated, and sometimes biased.

The phenology parameters for the deciduous model (p12–17) were reasonably estimated. While phenology provides a clear signal in the NEE observations (Fig. 4), there was limited information content on such discrete (i.e. on–off) processes in just two years of data. Continuous processes (i.e. operating every day), such as the temperature response of heterotrophic respiration (E_t), could make use of all available data and were well constrained. The poor estimate of parameter T_d, which controls the decomposition of litter to SOM, was likely due to the small relative size of this flux, the large magnitude differences in pool sizes of litter and SOM, and the independent pathways of respiration from both these pools. The decomposition flux could not thus be well estimated by NEE data alone.

The inability of the data to strongly constrain the confidence intervals on the F_g parameter, even though the mean estimates were good, was an unexpected result, given that NEE is sensitive to this parameter. The poor constraint (d_2 metric) is likely connected to the broad confidence intervals observed for the allocation parameters of NPP to foliage (F_{nf}) and roots (F_{nrr}). The ultimate cause of the poor constraint on NPP and foliage/root allocation is likely connected to the weak constraint on dynamics of wood and fine root C stocks. It seems that a range of different F_g, F_{nf} and F_{nrr} parameters within the model were able to produce NEE predictions reasonably consistent with observations, by varying the C content in the wood and root pools (i.e. equifinality).

We expected that there would be a weaker constraint on parameters with strong correlations. But there was no strong agreement among algorithms on which parameters were correlated. For the few correlations that were in common among algorithms, there was no evidence that the correlated parameters were less well determined, or biased. For correlated parameters, we found that in some case both were well determined, in some cases neither, and in other cases one parameter was well constrained. We found similar results using the synthetic and observed data. The minor differences in correlation patterns between EC and SYN datasets were to be expected, as correlations are largely a function of the model structure. However, with the EC datasets we expected to find evidence of more divergence among algorithms and more correlation and covariance due to model error. But there was no evidence of such, perhaps because of sparsity of observations and/or large uncertainties in the data.

4.2. Flux and stock estimation

There was weak agreement among algorithms in estimations of 90% CI on NEE and its component fluxes, for all datasets. The differences in CI size were closely related to differences among algorithms in parameter confidence intervals. There were considerable differences in assessments among similar algorithms (e.g. Metropolis), suggesting that the subjective choices of convergence tests versus statistical tests, priors for the parameters, and likelihood function within the method were important determinants of CI. None of the algorithms consistently included within the 90% confidence interval of the best-fit NEE ~90% of the synthetic true daily NEE values, or observed daily NEE data from year 3, (Tables 8 and 9). All algorithms at some point over- or underestimated the confidence interval. For annual assessments of NEE, GPP and R_e, there was more success, with some algorithms successful locating the true value from synthetic experiments within relatively narrow 90% confidence intervals (Fig. 5).

Assimilation results for annual flux predictions were in overall agreement with previous estimates from gap-filling studies on half-hourly data (Desai et al., 2008). However, in a number of cases the mean 90% CI did not include the gap-filled value (Fig. 6), for instance NEE in 2001 for Loobos. Some differences were to be expected, because the REFLEX database used only a subset of the measured data (when > 90% of half-hourly periods were measured in a day), and the assimilation was based on daily sums rather than half-hourly measurements. The general agreement in the partitioning of NEE into GPP and R_e using daily NEE data by REFLEX and half-hourly data by Desai et al. (2008) is notable. Respiration data can be easily extracted from hourly exchange data, but partitioning using daily data requires an effective GPP model, and sound predictions of foliar C. The partitioning result suggests that the DALEC GPP and phenology sub-models have worked reasonably at the FLUXNET sites. These results indicate that daily data are effective for model calibration, and that hourly resolution is not necessarily an advantage in generating predictions of annual C exchanges. We acknowledge that finer scale temporal data could provide extra constraint and perhaps reduce uncertainty, and this possibility needs further investigation (Sacks et al., 2006).

4.3. Model error

The magnitude of model error is difficult to calculate and rarely quoted. Because of the design of REFLEX, using both synthetic and observed data, we can produce a model error estimate by computing how much confidence intervals expand when using an uncertain model (in the case EC observations) versus a certain model (for SYN data). A comparison of confidence interval size on annual NEE estimates generated from synthetic and observed data revealed a common pattern, with larger CI for EC datasets (Table 10). Based on the comparison between CI on SYN and EC datasets, we conclude that the impact of model error was to increase the size of confidence intervals on annual NEE estimates by ~31%.

4.4. Prediction error

Prediction error, determined by forcing the model for 12 months beyond the assimilation period, was more complex to determine, because confidence intervals varied strongly between years 1 and 2 of the analysis. The only factor in common to all datasets was the lack of priors for initial conditions of C_f, C_{lit} and C_r. Thus, it is likely that erroneous initial conditions and/or large uncertainties on the initial values caused larger CI in year 1. The initial pools were often out of equilibrium with parameters, and so changed relatively quickly at first. By year two, parameter and state equilibria for these fast C pools reduced uncertainty. For predictions in year 3, lacking constraint of observations, uncertainty increased. CI on predictions (year 3) were > twice those for year 2 analyses. For the SYN experiments, the year 3 predictions among algorithms were similarly successful to years 1 and 2—that is, a similar fraction of 90% confidence intervals on annual flux estimates encompassed the truth. This result suggests that the quantification of increasing CI was reasonable.

4.5. Algorithm assessment

We examined the different algorithms, to determine if there were distinct winners or losers. All approaches produced broadly similar parameter retrievals (Fig. 2) for both synthetic and observed datasets (Tables 6 and 7). All approaches generated effective best estimates and predictions of daily NEE, as shown by the small RMSEs. But the focus of this study was also on the generation of sound confidence intervals to supplement these estimates. At the daily time-step the results were equivocal, with a tendency for algorithms to be over- or under-confident (Table 8). But at the annual timescale, perhaps the most relevant for C studies, we found that most of the algorithms' estimates of NEE encompassed the truth within 90% CI. A complementary test was to check the mean size of confidence intervals, to identify and weed out those cases where a successful test was obtained by using very broad CI. Thus, the test of annual NEE, GPP and R_e retrievals (year 1 and 2) and predictions (year 3) against the known truth from the synthetic experiments (Fig. 5) is perhaps the most useful judgement on the individual algorithms. According to this test, Metropolis methods, Kalman filters and genetic algorithms were all capable of correctly identifying a large proportion of true fluxes with relatively small confidence intervals. Thus all approaches were valid, but some implementations were more effective in terms of this test on confidence intervals than others (see Appendix A for more information on algorithms). The sequential updating of the Kalman filters, allowing shifts in states through the model run unconnected to parameters, may be connected to the success of such methods (E1, E2, U1) in generating effective, but narrow confidence intervals.

Some algorithms (M4 and M5) had problems with large changes to C_w and C_{som} pools, which could be made consistent with the flux data, but are not ecologically sound in an undisturbed ecosystem. This problem seems to be partly related to a steady state assumption, with stocks first confined to an equilibrium, which likely leads to an erroneous initial system state, potential biases in parameters and inflation of their confidence intervals, as shown recently in a specific study by Carvalhais et al. (2008). These symptoms are, for example, also seen in the approach M4, where a spin-up was performed. Hence, a way to estimate the initial state of the system without an *ad-hoc* steady state assumption is crucial to successful MDF and should be explored further. A constraint on the annual changes in these pools based on repeated inventories would help solve this problem. Stem inventories are likely to be easier to undertake with quantifiable error than those on SOM, and so should be the focus for future studies. Nevertheless, if longer time scale are to be addressed there is a need to imposed constraints from soil carbon data, e.g. via chronosequences or profile data. Some algorithms did not explicitly include searching for initial conditions on C_f, C_{lab} and C_{lit}, and this caused some problems for e.g. E2. All algorithms need to assess their estimates of uncertainties and develop new approaches for uncertainty estimates that are consistent with the observations.

This experiment has demonstrated the value of using synthetic datasets in understanding data assimilation problems. It is clear that even with a perfect model, existing model-data fusion approaches find it difficult to analyse parameters using synthetic, noisy and sparse datasets. The information content of data that can be extracted by MDF depends on data quality and coverage. Further synthetic studies will illuminate the relationship between data availability and parameter constraint. It is clear that there is little consensus on how to generate confidence intervals, with very broad ranges among algorithms. Tests using confidence intervals provide a useful first look at assessing the uncertainties quantified by the various algorithms, although representing continuous probability distributions with a confidence interval suffers from using an arbitrary cutoff criteria. Algorithms that are not well constrained by the data, and thus have wide CI's, will be more likely to contain the true value but this suggests they are less able to make use of all the information in the data.

5. Conclusions

A range of model-data fusion algorithms exist that can generate useful estimates of parameter probability density functions and state estimates for C models using daily net ecosystem exchange data, derived either from observations or synthetically. While there was less agreement among algorithms on the size of confidence intervals on parameter and state estimates, some algorithms were able to make effective estimates of annual fluxes within relatively small CI, when compared to detailed gap-filled estimates or the synthetic 'truth'. Overall, algorithms generated narrower confidence intervals in analyses using synthetic data compared to observed data. Likewise, confidence intervals were larger by 88% for forecast periods than during data-fusion periods. These results suggest that some algorithms were generally able to make a reasonable quantification of error propagation in prediction periods, and of the likely size of model error, but that differences in estimated confidence intervals suggests further improvements are required. Further studies should explore the importance of assumptions about parameter priors (Gaussian or uniform), and the handling of unknown initial conditions. Exploring the growth in CI over forecast periods of multiple years also needs to be explored in a further study. Seasonal data on the variation in slow large C pools would be a useful addition to model-data fusion studies, even with large confidence intervals. Such data can help constrain the parameters poorly served by eddy covariance data, which are those related to allocation of photosynthate to respiration and plant pools, and turnover of wood and roots.

Acknowledgements

NERC funded much of this work through the CarbonFusion International Opportunities grant. MW, CT, AF, MR and AR were involved in the initial planning and discussions for REFLEX at the CarbonFusion meeting in Edinburgh, May 2006. AF undertook the main work on setting up the experiment, collating results, and undertaking analyses. MW initiated and managed the experiment, and led the production of the manuscript. AR was involved in defining the analytical process and the parameter estimation assessment. Other authors participated in the experiment and all contributed to the paper. We are grateful to A Granier and the Hesse research team, and to EJ Moors and the Loobos research team, for access to their data, and the FLUXNET team for data processing and preparation. We are grateful to Mike Raupach and Damian Barrett for their ideas and input at the start of the project, and in developing the protocol for the experiment outlined here. We also recognise Zhang Li's efforts and input to the analysis. Jens Kattge made useful comments on the manuscript. AR acknowledges support from the Office of Science (BER), U.S. Department of Energy, through the Northeastern Regional Center of the National Institute for Climatic Change Research. We thank two anonymous reviewers for their comments.

Appendix A

A.1. Details of algorithms

E1: Combined Metropolis–Ensemble Kalman Filter (two–stage approach). Stage 1. Parameter estimation. Parameters were initially estimated using a simple Metropolis MCMC-type approach. (e.g. (Knorr and Kattge, 2005; Mosegaard and Tarantola, 1995)). Initial prior distributions were assumed to be uniform and encompass the entire possible suggested range and so a single stage accept/reject criterion was used based on comparison of model output with data alone. Initial values for C_r, C_{lit} and C_{lab} were estimated in the same manner as parameters, initial values for C_f were based on first available observation (EV case) or set to zero (DE case). The model was initialized from a random location, and step size was constant and determined as 0.001 of log-transformed parameter range. This was determined through 'tuning' initial runs to ensure an acceptance probability of between 0.2 and 0.8 at each step. The number of steps required to sufficiently sample the parameter space was assessed using the Gelman criteria (Gelman, 1995) to test convergence between chains.

E1: Stage 2. State estimation. Eight thousand parameter sets were randomly sampled from the accepted parameters from Stage 1. These were then used in an 8000 member Ensemble Kalman Filter (EnKF, Evensen, 1994; Williams et al., 2005). A unique parameter set was assigned to each ensemble member with the intention this would cause divergence between ensemble members representing model error and cause a growth in state uncertainty equivalent to that inherent from parameter uncertainty alone. This was done instead of adding a stochastic forcing term at each time-step. This is possibly correct in the SYN cases when model 'structural' error is known to be zero, but will probably underestimate model error in the EC cases and overly restrict growth in state uncertainty. Nonetheless, assimilation of observations did alter the state variables in the resulting analysis and reduce uncertainties in state estimates even though these same observation data had already been used to generate the parameter sets in Stage 1 so offered little additional information to the EnKF.

E2: Ensemble Kalman Filter. This method was set up for joint estimation of states and parameters, so parameters were included within the state vector for assimilation. Model parameter errors were set within bounds—small enough to avoid tracking daily noise in observations, and large enough to shift over weekly-seasonal timescales in response to process signals. Errors on model states were set smaller than for parameters, so that assimilation was focused on updating parameters rather than states. Initial values for all parameters and initial conditions for $C_{f,}$ C_{lab} and C_r were estimated. After an initial assimilation of observations, these initial parameter estimates were updated with the final estimates from the assimilation. We assumed that $C_{f,}$ C_{lab} and C_r would be in approximate steady state over annual cycles, and adjusted initial values accordingly. A further EnKF assimilation was then applied, using these updated initial parameters and initial conditions, to generate final analyses.

U1: Combined Metropolis–Unscented Kalman Filter. The UKF was used to provide state estimates for each of the experiments. The UKF (Julier and Uhlmann, 2004) is a nonlinear version of the traditional linear Kalman filter (Kalman, 1960), that uses a deterministic sampling of so-called sigma points in order to capture the mean and covariance of the state. Similar to other Kalman type filters it employs a two-step 'predictor–corrector' scheme where model predictions are corrected by measurements as they arrive sequentially in time. At time periods where measurements are missing, only the prediction step is used. To employ the UKF, the general nonlinear state space model was assumed, with the variants of the model taking the form of the state evolution equations. A linear measurement model was used in all runs. Both the state and measurement equations assume zero mean random noise processes with associated full-dimensional covariance matrices (Gove and Hollinger, 2006). The later were estimated from the information provided. The parameter estimates used in the filter runs were arrived at via simulated annealing method M3. Parameters for the unscented transformation were set to $\alpha = 1$, $\beta = 2$ and $\kappa = 1$ for all experiments (see Gove and Hollinger, 2006 for an explanation).

G1: Genetic algorithm: The implementation was from Haupt and Haupt (2004). The population size was 100, and was run for 1000 iterations (generations). Initial stores Cr, C_{lit}, C_{f} and C_{lab} were estimated by the GA as additional parameters. To estimate uncertainties, the roughly 1600 (unique) parameter sets with cost function values closest to the final best cost function value were saved, and used to estimate the covariance matrix and 90% CI.

M1: Metropolis. This method sought to make as few approximations as possible to Bayes Theorem, choosing the simplest algorithm to generate a representative sample from the posterior. We chose the beta distribution for our prior. The Metropolis algorithm (Metropolis et al. 1953) generates a chain that sequentially "walks through parameter space" in such a way that the chain of visited points is the sought-after sample from the posterior. Each new point in the chain is found by randomly generating a multivariate normal step away from the current vector. In this case a simple diagonal variance matrix defined this multivariate normal "proposal distribution". Whether a proposed candidate vector was accepted or not depended on the *Metropolis ratio*, which is the ratio of two products: likelihood times prior for the candidate and likelihood times prior for the current point. If the Metropolis ratio was larger than 1 (i.e. the candidate point has a higher posterior probability then the current point), it was always accepted. If the Metropolis ratio was less than 1 (i.e. the candidate was "less probable" than the current vector), the candidate could still be accepted but only with probability equal to the Metropolis ratio. The chain was stopped when it "converged", i.e. it had explored the parameter space adequately. Convergence was confirmed visually using the trace plots of the different parameters, i.e. plots that show how the chain moves through parameter space for each individual parameter. If one or more of the trace plots was still showing drift towards unexplored parts or parameter space, the chain was deemed not to have converged.

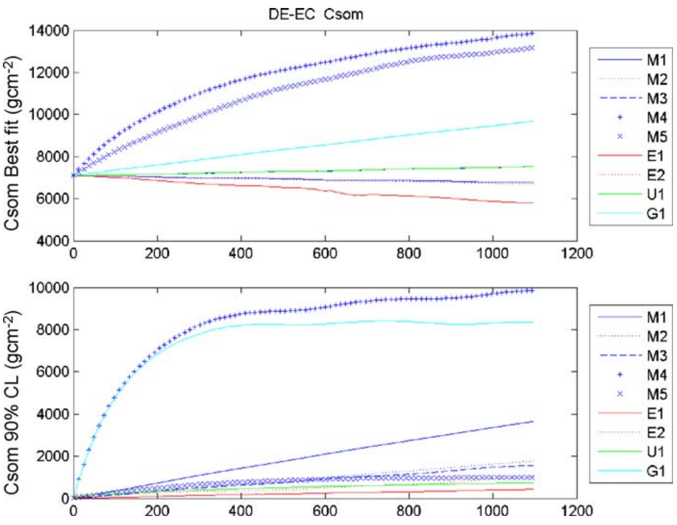
M2: Combined genetic algorithm–Metropolis (2 stage process). A combined optimisation approach estimated model parameters and state variables. A genetic algorithm, Stochastic Evolutionary Ranking Strategy (SRES) was used to find the global optimum (Runarrsson and Yao, 2000). Markov chain monte carlo (MCMC) using the Metropolis–Hastings algorithm was then used to explore the parameter space around the optimum to estimate the full joint distribution of parameters and to estimate predictive uncertainty. Two chains were run for each experiment; convergence was determined by visually comparing the parameter PDFs from both chains. The ranges given for p1-17 were used as uniform distributions; no additional information was used. The initial values of pools C_{r}, C_{lit} and C_{lab} were also estimated as model parameters, using the prior range 20–200 gC m^{-2} as recommended. All observations are assumed to drawn from independent distributions. Both NEE and LAI errors were assumed normally distributed.

M3: Metropolis. Optimisation of parameters and initial values of C pools took place in three stages. First, the parameter and initial state space was randomly explored for 50,000 iterations, at which point the parameter set and initial conditions with the lowest cost function was used as the starting point for the Metropolis algorithm. Second, the Metropolis algorithm was implemented to ensure progressive downslope movement while at the same time avoiding local minima. The cost function was a weighted-sum-of-squares of both NEE and LAI deviations. 200,000 steps were taken in this manner. Third, reverting to the best parameter set obtained, the parameter space was explored again until 1000 parameter sets have been accepted as "almost as good as" the optimal parameter set, using a χ^2 test to determine the threshold contour (90% confidence interval) (assuming $n-1$ degrees of freedom for LAI and $n-p-1$ degrees of freedom for NEE. These parameter sets were used to define the uncertainty estimates on both parameters and model predictions.

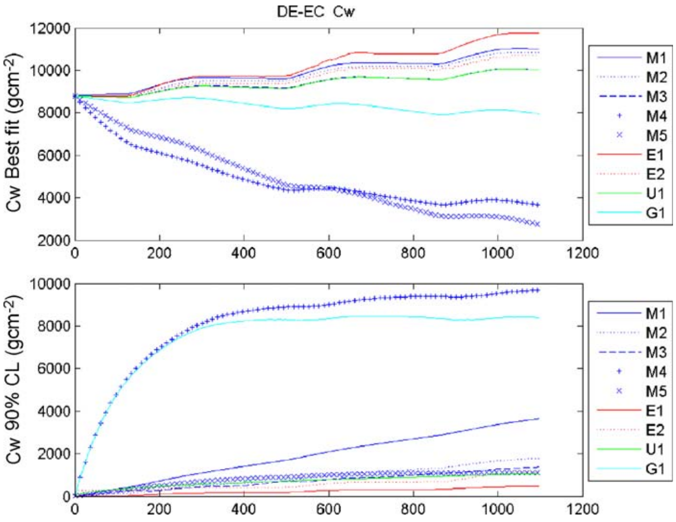
M4: Metropolis. The algorithm adopted a global search method with an uniform walk in the model parameter space. The method is based on a Bayesian approach where the comparison between model output and data is used to update our prior knowledge of the parameter distribution. The prior distributions were considered to be uniform. The Metropolis rules prevented the algorithm from being trapped in local minima, allowing for changes in the searching direction. Spin-up was used to initialize the C pools (C_{r}, C_{lit} and C_{lab}); we sampled the parameters and we ran the model replicating the meteorological data until the total difference between one year and the other was less than 1 g of C. The other C pools were initialized as from the experiment description.

M5. Metropolis. The SCEM-UA algorithm (Vrugt et al., 2003) is a modified version of the original SCE-UA global optimisation algorithm (Duan et al., 1992). The algorithm is Bayesian in nature and operates by merging the strengths of the Metropolis algorithm, controlled random search, competitive evolution, and complex shuffling to continuously update the proposal distribution and evolve the sampler to the posterior target distribution. The SCEM-UA algorithm uses the Metropolis–Hastings (Metropolis et al., 1953) search strategy to generate a sequence of parameter sets (θ_1, θ_2, . . ., θ_n) that adapts to the target posterior distribution. It starts with an initial population of points (parameter sets) randomly distributed throughout the feasible parameter space defined by the prior parameter distributions. The population is partitioned into q complexes, and in each complex k ($k = 1, 2, \ldots, q$) a parallel sequence is launched from the point that exhibits the highest posterior density. A new candidate point in each sequence k is generated using a multivariate normal distribution either centred around the current draw of the sequence k, or the mean of the points in complex k, augmented with the covariance structure induced between the points in complex k. The Metropolis-annealing criterion is used to test whether the candidate point should be added to the current sequence. Subsequently the new candidate point randomly replaces an existing member of the complex. Finally, after a certain number of iterations new complexes are formed through a process of shuffling the old complexes. The objective function used in this study is a combination of the model errors (expressed as SSE, Sum of Squared Errors) of describing the CO_2 fluxes and the Leaf Area Index, weighted by the error variance of each variable.

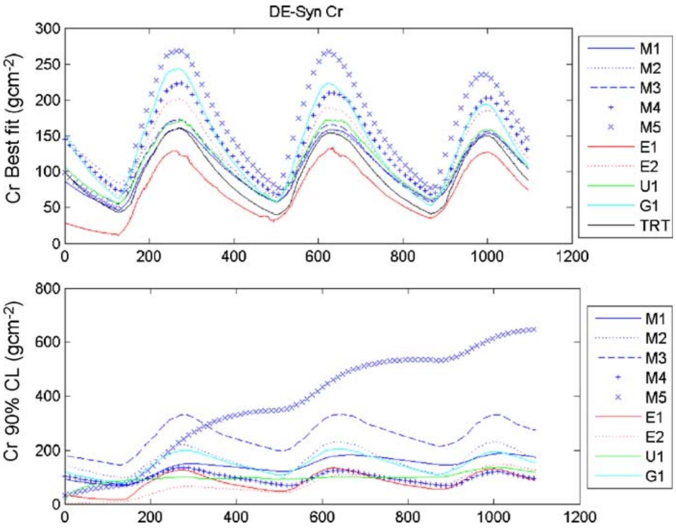
A.2. Detailed model outputs



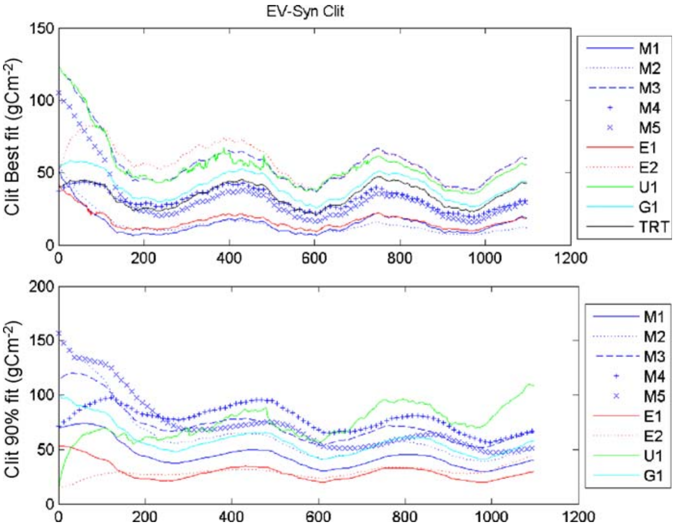
Retrieved estimates of soil organic matter/coarse woody debris C stocks over three years for the DE-EC deciduous site with observations of NEE fluxes and LAI assimilated. The upper panel shows best-fit or mean for C_{som}, and the lower panel shows the width of the 90% confidence interval. Algorithms are indicated by the codes in the right hand panels.



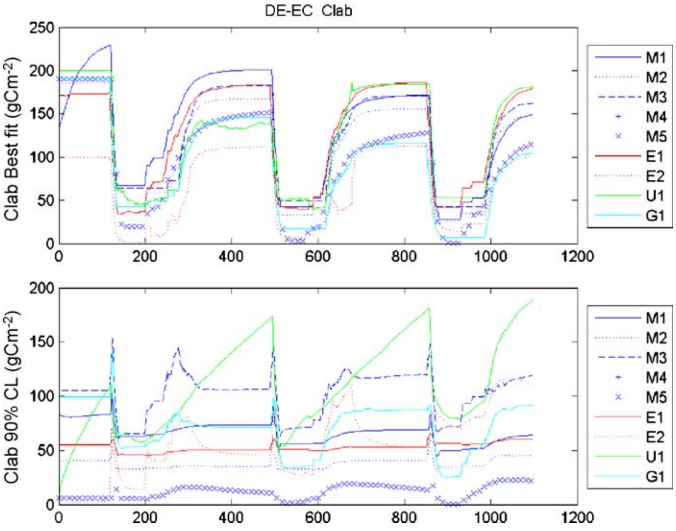
Retrieved estimates of woody C stocks over three years for the DE-EC deciduous site with observations of NEE fluxes and LAI assimilated. The upper panel shows best-fit or mean for C_w, and the lower panel shows the width of the 90% confidence interval. Algorithms are indicated by the codes in the right hand panels.



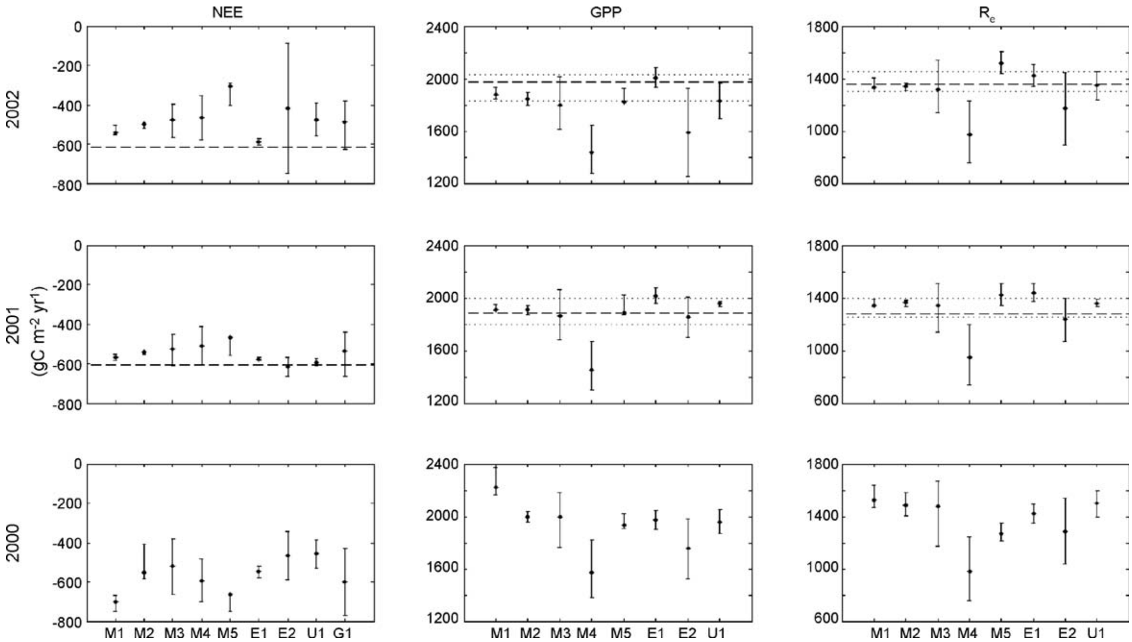
Retrieved estimates of fine root C stocks over three years for the DE-SYN deciduous site with synthetic NEE fluxes and LAI assimilated. The upper panel shows best-fit or mean for C_r, and the lower panel shows the width of the 90% confidence interval. The true value of C_r is indicated in the upper panel also. Algorithms are indicated by the codes in the right hand panels.



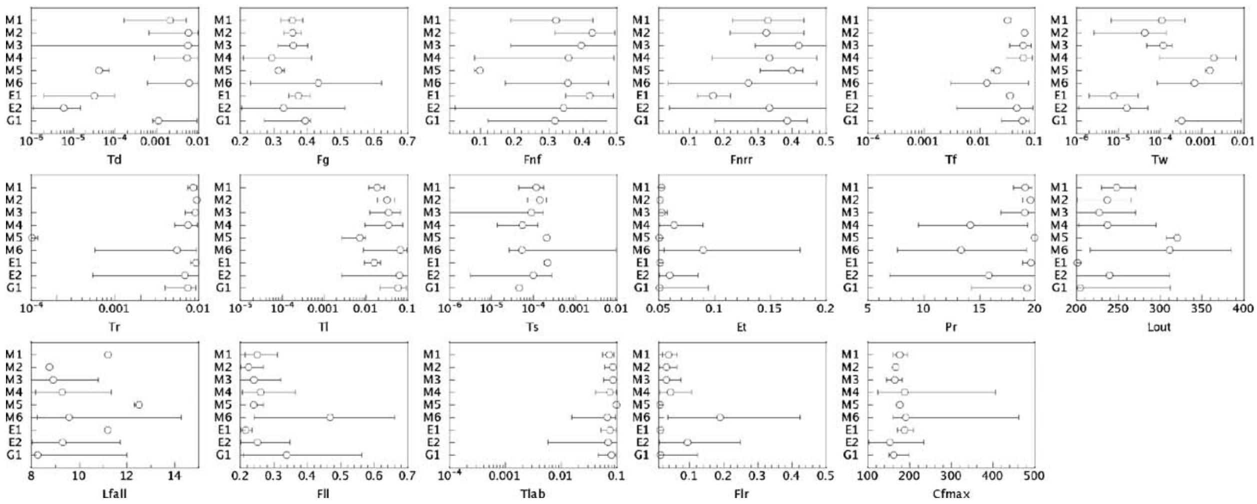
Retrieved estimates of litter C stocks over three years for the EV-SYN evergreen site with synthetic NEE fluxes and LAI assimilated. The upper panel shows best-fit or mean for C_{lit}, and the lower panel shows the width of the 90% confidence interval. The true value of C_{lit} is indicated in the upper panel also. Algorithms are indicated by the codes in the right hand panels.



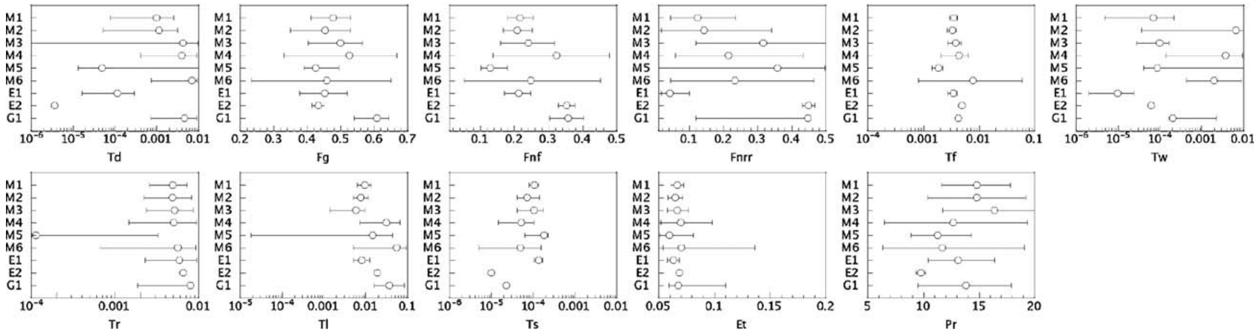
Retrieved estimates of labile C stocks over three years for the DE-EC deciduous site with observed NEE fluxes and LAI assimilated. The upper panel shows best-fit or mean for C_{lab}, and the lower panel shows the width of the 90% confidence interval. Algorithms are indicated by the codes in the right hand panels.



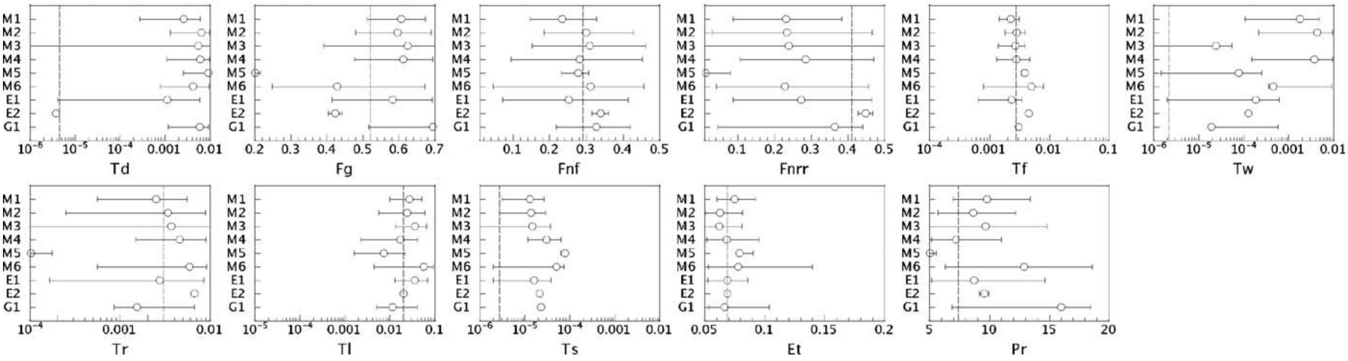
Annual analyses of NEE, GPP and R_e for 2000, 2001 and prognoses for 2002 generated with the DE-EC dataset from Hesse, France. Results are shown for each algorithm for NEE and for eight algorithms for GPP and R_e, with 90% confidence intervals indicated. The dashed lines show the best estimates from gap-filling routines using hourly NEE data, while the dotted lines show interquartile range among the estimates from the array of gap-filling routines for 2001 and 2002 (Desai et al., 2008).



Parameter estimation for deciduous FLUXNET (DE-EC) data. The panels shows each of the algorithms' best estimate of each parameter, and the magnitude of each 90% confidence intervals. The upper and lower bounds of each parameter, as provided to the experimenters, is indicated by the range of each x-axis. x-Axes are log scaled for turnover rates (all parameters beginning T). For an explanation of parameter symbols see Table 5.



Parameter estimation for evergreen FLUXNET (EV-EC) data. The panels shows each of the algorithms' best estimate of each parameter, and the magnitude of each 90% confidence intervals. The upper and lower bounds of each parameter, as provided to the experimenters, is indicated by the range of each *x*-axis. *x*-Axes are log scaled for turnover rates (all parameters beginning *T*). For an explanation of parameter symbols see Table 5.



Parameter estimation for evergreen synthetic (EV-SYN) data. The panels shows each of the algorithms' best estimate of each parameter, and the magnitude of each 90% confidence intervals. The 'true' value of the parameter used in generating the synthetic data is indicated by the *d* vertical line. The upper and lower bounds of each parameter, as provided to the experimenters, is indicated by the range of each *x*-axis. *x*-Axes are log scaled for turnover rates (all parameters beginning *T*). For an explanation of parameter symbols see Table 5.

References

Baldocchi, D., Falge, E., Gu, L.H., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X.H., Malhi, Y., Meyers, T., Munger, W., Oechel, W., Paw U, K.T., Pilegaard, K., Schmid, H.P., Valentini, R., Verma, S., Vesala, T., Wilson, K., Wofsy, S., 2001. FLUXNET: a new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bulletin of the American Meteorological Society 82 (11), 2415–2434.

Bonan, G.B., 2008. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. Science 320 (5882), 1444–1449.

Braswell, B.H., Sacks, W.J., Linder, E., Schimel, D.S., 2005. Estimating diurnal to annual ecosystem parameters by synthesis of a carbon flux model with eddy covariance net ecosystem exchange observations. Global Change Biology 11, 335–355.

Carvalhais, N., Reichstein, M., Seixas, J., James Collatz, G., Santos Pereira, J., Berbigier, P., Carrara, A., Granier, A., Montagnani, L., Papale, D., Rambal, S., Sanz, M.J., Valentini, R., 2008. Implications of carbon cycle steady state assumptions for biogeochemical modeling performance and inverse parameter retrieval. Global Biogeochemical Cycles GB2007.

Davidson, E.A., Janssens, I.A., 2006. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. Nature 440, 04514.

Desai, A.R., Richardson, A.D., Moffat, A.M., Kattge, J., Hollinger, D.Y., Barr, A., Falge, E., Noormets, A., Papale, D., Reichstein, M., Stauch, V.J., 2008. Cross site evaluation of eddy covariance GPP and RE decomposition techniques. Agricultural and Forest Meteorology 148, 821–838.

Dewar, R.C., Franklin, O., Makela, A., McMurtrie, R.E., Valentine, H.T., 2009. Optimal function explains forest responses to global change. Bioscience 59 (2), 127–139.

Duan, Q., Sorooshian, S., Gupta, V.K., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resources Research 28, 1015–1031.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. 99, 10143–10162.

Evensen, G., 2003. The Ensemble Kalman Filter: Theoretical Formulation and Practical Implementation. Ocean Dynamics 53, 343–367.

Farquhar, G.D., von Caemmerer, S., 1982. Modelling of photosynthetic response to the environment. In: Lange, O.L., Nobel, P.S., Osmond, C.B., Ziegler, H. (Eds.), Physiological Plant Ecology II. Encyclopedia of Plant Physiology, New Series. Vol. 12B. Encyclopedia of Plant Physiology. Springer-Verlag, Berlin, pp. 549–587.

Gelman, A., 1995. Bayesian data analysis. In: Texts in Statistical Science, Chapman & Hall, London, xix, 526 pp.

Gelman, A., Rubin, D.B., 1992. Markov chain Monte Carlo methods in biostatistics. Statistical methods in Medical Research 5, 339–355.

Gove, J.H., Hollinger, D.Y., 2006. Application of a dual unscented Kalman filter for simultaneous state and parameter estimation in problems of surface-atmosphere exchange. Journal of Geophysical Research 111 D08S07.

Julier, S.J., Uhlmann, J.K., 2004. Unscented filtering and nonlinear estimation. Proceedings of the IEEE 92, 410–422.

Haupt, R.L., Haupt, S.E., 2004. Practical Genetic Algorithms. Wiley Blackwell. 272 pp.

Heidelberger, P., Welch, P.D., 1983. Simulation run length control in the presence of an initial transient. Operations Research 31, 1109–1144.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Transactions of the ASME – Journal of Basic Engineering 82, 35–45.

Knorr, W., Kattge, J., 2005. Inversion of terrestrial ecosystem model parameter values against eddy covariance measurements by Monte Carlo sampling. Global Change Biology 11 (8), 1333–1351.

Lasslop, G., Reichstein, M., Kattge, J., Papale, D., 2008. Influences of observation errors in eddy flux data on inverse model parameter estimation. Biogeosciences 5, 1311–1324.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1092.

Mosegaard, K., Tarantola, A., 1995. Monte-Carlo sampling of solutions to inverse problems. Journal of Geophysical Research-Solid Earth 100 (B7), 12431–12447.

Raupach, M.R., Rayner, P.J., Barrett, D.J., DeFries, R.S., Heimann, M., Ojima, D.S., Quegan, S., Schmullius, C.C., 2005. Model-data synthesis in terrestrial carbon observation: methods, data requirements and data uncertainty specifications. Global Change Biology 11, 378–397.

Richardson, A.D., Hollinger, D.Y., 2005. Statistical modeling of ecosystem respiration using eddy covariance data: maximum likelihood parameter estimation, and Monte Carlo simulation of model and parameter uncertainty, applied to three simple models. Agricultural and Forest Meteorology 131, 191–208.

Richardson, A.D., Mahecha, M.D., Falge, E., Kattge, J., Moffat, A.M., Papale, D., Reichstein, M., Stauch, V.J., Braswell, B.H., Churkina, G., Kruijt, B., Hollinger, D.Y., 2008. Statistical properties of random CO_2 flux measurement uncertainty inferred from model residuals. Agricultural and Forest Meteorology 148, 38–50.

Sacks, W.J., Schimel, D.S., Monson, R.K., Braswell, R.H., 2006. Model-data synthesis of diurnal and seasonal CO_2 fluxes at Niwot Ridge, Colorado. Global Change Biology 12, 240–259.

Schulz, K., Jarvis, A., Beven, K., Soegaard, H., 2001. The predictive uncertainty of land surface fluxes in response to increasing ambient carbon dioxide. Journal of Climate 14, 2551–2562.

Sitch, S., Huntingford, C., Gedney, N., Levy, P.E., Lomas, M., Piao, S.L., Betts, R., Ciais, P., Cox, P., Friedlingstein, P., Jones, C.D., Prentice, I.C., Woodward, F.I., 2008. Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five Dynamic Global Vegetation Models (DGVMs). Global Change Biology 14 (9), 2015–2039.

Trudinger, C.M., Raupach, M.R., Rayner, P.J., Kattge, J., Liu, Q., Pak, B., Reichstein, M., Renzullo, L., Richardson, A.D., Roxburgh, S.H., Styles, J., Wang, Y.P., Briggs, P., Barrett, D., Nikolova, S., 2007. OptIC project: an intercomparison of optimization techniques for parameter estimation in terrestrial biogeochemical models. Journal of Geophysical Research-Biogeosciences 112 (G2), G02027.

Van Wijk, M.T., Williams, M., Laundre, J.A., Shaver, G.R., 2003. Interannual variability of plant phenology in tussock tundra: modelling interactions of plant productivity, plant phenology, snow melt and soil thaw. Global Change Biology 9, 743–758.

Vrugt, J.A., Gupta, H.V., Bouten, W., Sorooshian, S., 2003. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resources Research 39 (8) Art. no. 1201.

Wang, Y.P., Baldocchi, D., Leuning, R., Falge, E., Vesala, T., 2007. Estimating parameters in a land-surface model by applying nonlinear inversion to eddy covariance flux measurements from eight FLUXNET sites. Global Change Biology 13, 652–670.

Waring, R.H., Landsberg, J.J., Williams, M., 1998. Net primary production of forests: a constant fraction of gross primary production? Tree Physiology 18, 129–134.

Williams, M., Rastetter, E.B., Fernandes, D.N., Goulden, M.L., Shaver, G.R., Johnson, L.C., 1997. Predicting gross primary productivity in terrestrial ecosystems. Ecological Applications 7 (3), 882–894.

Williams, M., Rastetter, E.B., Fernandes, D.N., Goulden, M.L., Wofsy, S.C., Shaver, G.R., Melillo, J.M., Munger, J.W., Fan, S.-M., Nadelhoffer, K.J., 1996. Modelling the soil–plant–atmosphere continuum in a *Quercus-Acer* stand at Harvard Forest: the regulation of stomatal conductance by light, nitrogen and soil/plant hydraulic properties. Plant, Cell and Environment 19, 911–927.

Williams, M., Schwarz, P., Law, B.E., Irvine, J., Kurpius, M.R., 2005. An improved analysis of forest carbon dynamics using data assimilation. Global Change Biology 11, 89–105.

Zhang, Y.J., Xu, M., Chen, H., Adams, J., 2009. Global pattern of NPP to GPP ratio derived from MODIS data: effects of ecosystem type, geographical location and climate. Global Ecology and Biogeography 18 (3), 280–290.